Recherches sur le déflationnisme aléthique contemporain

Julien Boyer

Université Paris 1 Panthéon-Sorbonne
ED 280 - UFR de Philosophie

IHPST

Thèse

pour l'obtention du grade de docteur en Philosophie
de l'Université Paris 1 Panthéon-Sorbonne

présentée et soutenue publiquement
le 16 décembre 2019 par

Julien Boyer

Recherches sur le déflationnisme aléthique contemporain

Sous la direction de Pierre Wagner

Composition du Jury :

Mme. Hourya Benis	Directrice de Recherche au CNRS
M. Henri Galinon	Maître de conférences à l'Université Clermont Auvergne
M. Leon Horsten	Professeur à l'Université de Constance
M. Philippe de Rouilhan	Directeur de Recherche au CNRS
M. Gabriel Sandu	Professeur à l'Université d'Helsinki
M. Pierre Wagner	Professeur à l'Université Paris 1 Panthéon-Sorbonne

Année 2019

Julien Boyer

Recherches sur le déflationnisme aléthique contemporain

Thèse

pour l'obtention du grade de docteur en Philosophie
de l'Université Paris 1 Panthéon-Sorbonne

Sous la direction
de Pierre Wagner

À la mémoire d'Anne-Marie Boyer
dont l'amour et le souvenir demeurent

Remerciements

En premier lieu, je tiens à remercier mon directeur de thèse, Pierre Wagner, d'avoir accepté d'encadrer mon travail et de m'avoir permis de le mener à terme. Ses relectures patientes, ses innombrables commentaires, suggestions et corrections ont grandement amélioré la qualité de ce manuscrit. Qu'il en soit chaudement remercié ici.

Je remercie également Hourya Sinaceur, Henri Galinon, Leon Horsten, Philippe de Rouilhan et Gabriel Sandu d'avoir accepté de participer à mon jury de thèse.

J'adresse aussi un merci supplémentaire à Gabriel Sandu pour avoir accompagné mes premiers pas en logique et en philosophie formelle et avoir, le premier, suscité mon intérêt pour le déflationnisme en matière de vérité.

Merci à Marco Panza pour m'avoir accompagné et soutenu un temps lors du développement de ce travail.

Au cours de mes années de formation à l'IHPST, j'ai pu interagir et discuter avec de nombreux chercheurs et ainsi bénéficier de leurs conseils. Certains d'entre eux ont accepté de lire ou d'assister à des présentations de ce travail dans des versions antérieures. Outre Pierre Wagner, je pense à Philippe de Rouilhan, Marco Panza, Fabrice Pataut, Gabriel Sandu, Susana Berestovoy et Henri Galinon. Merci à eux pour leurs remarques et suggestions qui m'ont permis d'améliorer grandement mes arguments et d'approfondir mon point de vue. Merci également à Denis Bonnay pour les stimulantes discussions que nous avons pu avoir sur le déflationnisme ou sur d'autres sujets touchant la logique et la philosophie. Merci à Ekaterina Kubyshkina et à Victor Lefèvre pour leur investissement au sein de l'IHPST. Merci à Méven Cadet pour nos amicales discussions à la BNF ou ailleurs. Merci à Marion Vorms et Henri Galinon pour les bons moments passés dans les locaux du DEC lorsque nous y partagions un bureau.

Je voudrais ensuite adresser un merci particulier à Anouk Barberousse et Paul Egré qui m'ont tous deux tendu la main lorsque je traversais des moments difficiles au cours

de ces années d'étudiant et que le découragement me gagnait. J'ai été très touché par leur disponibilité et leur gentillesse désintéressées. Merci à eux deux.

Enfin, je remercie ma famille, à laquelle ce travail est dédié. Merci à mes parents, Anne-Marie et Jean-Claude Boyer qui m'ont toujours soutenu et encouragé à suivre ma propre voie. Merci à ma sœur, Jeanne, indispensable présence dans les bons et les moins bons moments. Merci à Elisabeth, ma compagne, dont l'amour, la patience et le soutien ont été si précieux. Merci à elle, à Anouk et à Etienne pour la joie et le bonheur qu'ils m'apportent jour après jour.

Sommaire

Introduction

Qu'est-ce que la Vérité ? (Pilate)
La vérité est décitation. (Quine)

À la question : « qu'est-ce-que la vérité ? », les réponses des philosophes ne manquent pas. Il y aurait plutôt surabondance de biens en la matière. De prime abord, il semble évidemment banal d'affirmer que la notion de vérité est centrale en philosophie. Elle est présente dans bien des domaines et paraît jouer un rôle primordial en épistémologie et en philosophie de la connaissance —bien sûr, mais aussi en philosophie des sciences, en philosophie du langage, en logique —sans doute, et peut-être également en métaphysique, voire en éthique ou en philosophie de l'action, *etc.* La caractérisation et la compréhension de cette notion est par conséquent une lourde et impérieuse tâche qui a occupé les philosophes depuis l'Antiquité. Parmi les conceptions traditionnelles [1], on peut distinguer diverses théories qui ont été proposées. Il y a d'abord la vénérable théorie de la vérité-correspondance selon laquelle un énoncé ou une proposition sont vrais si et seulement si ils correspondent à la réalité ou à un fait. D'après la théorie cohérentiste de la vérité, au contraire, un énoncé sera déclaré vrai s'il est membre d'un certain ensemble cohérent de croyances. La théorie vérificationniste quant à elle identifie la vérité d'un énoncé avec sa prouvabilité ou sa capacité à être ultimement vérifié dans des conditions d'enquête idéales. À ces trois premières conceptions s'ajoute la théorie pragmatiste de la vérité selon laquelle est vrai ce qu'il est utile de croire, tandis que d'autres philosophes ont également proposé de voir dans la vérité une propriété primitive et inanalysable qu'un énoncé possède ou ne possède pas sans qu'il soit possible d'en donner une plus ample explication. Ces paradigmes concurrents illustrent bien l'absence de consensus parmi les

1. Nous ne mentionnons ces conceptions que pour illustrer le genre de théories auxquelles s'opposent les déflationnistes. Nous n'en entreprendrons pas ici l'étude détaillée. Nous renvoyons le lecteur intéressé aux entrées pertinentes de la Stanford Encyclopedia, ou encore à Candlish et Damnjanovic (2007).

philosophes concernant la nature profonde de la vérité. Toutefois, tous s'accordent à considérer la vérité comme une propriété importante dont l'élucidation doit permettre d'éclairer de larges pans de la philosophie et de notre entreprise de connaissance.

C'est en quelque sorte principalement à ce dernier présupposé que s'oppose le « déflationnisme en matière de vérité » ou « déflationnisme aléthique [2] » auquel notre travail est consacré. L'objet de la présente étude est en effet d'introduire, d'analyser et d'évaluer divers arguments récemment développés autour de ce type de théories. Au seuil d'un examen critique du déflationnisme en matière de vérité, les canons d'une bonne méthodologie commandent qu'on commence par définir précisément ce dont il s'agit. Mais la satisfaction de cette exigence se heurte à un obstacle dans le cas particulier du déflationnisme. En effet, le terme ne désigne pas une doctrine unique et pleinement articulée mais plutôt un faisceau de doctrines, une famille de pensée. Et derrière cet étendard se regroupent des auteurs aux conceptions variées et parfois significativement différentes. Malgré tout, nous pouvons dès à présent remarquer que le terme signale bien qu'il s'agit de « déflater » ou de « dégonfler » ce qu'on attend généralement d'une théorie de la vérité : les déflationnistes ont ceci en commun qu'ils proposent de revoir à la baisse les exigences qui accompagnent ordinairement l'analyse philosophique de la notion de vérité. Le déflationnisme en matière de vérité s'est donc constitué en grande partie par opposition et en réaction aux diverses conceptions traditionnelles de la vérité.

Malgré l'aura de mystère qui a souvent entouré ce concept, le déflationniste aléthique considère en effet que la notion de vérité ne possède pas une quelconque nature profonde qu'il appartiendrait au philosophe de dévoiler et d'expliquer. Le prédicat « vrai » désigne au contraire une propriété quasi-triviale dont il est vain d'attendre une analyse réductive explicitant en toute généralité ses conditions d'application. Au delà de cette caractérisation donnée en creux et par contraste avec les conceptions traditionnelles, nous pouvons également ajouter que les déflationnistes s'accordent généralement pour attribuer un rôle

2. Il peut paraître curieux d'introduire une thèse de doctorat français en employant ce qui ressemble fort à un double barbarisme. Toutefois, on trouve l'adjectif « aléthique » (du grec $\alpha\lambda\acute{\eta}\theta\epsilon\iota\alpha$, vérité) dans certains dictionnaires de philosophie (*cf.* par exemple *Encyclopédie Universalis* (2018), *Dictionnaire de français Larousse* (2018)) comme un terme technique de logique modale désignant les modalités de la vérité (le vrai/le faux, le possible/l'impossible, le nécessaire/le contingent). Pour des raisons stylistiques, nous l'emploierons ici en un sens relâché pour désigner « ce qui concerne la vérité », de façon à pouvoir éviter parfois les longues périphrases du type « en matière de vérité », à propos de la vérité », « touchant la vérité », « concernant la vérité » *etc.* Pour ce qui est de l'anglicisme « déflationnisme » (ou l'adjectif « déflationniste »), il est désormais entré dans l'usage philosophique francophone pour qualifier une certaine école de pensée issue de la tradition analytique anglo-saxonne. Nous l'utiliserons donc librement et sans guillemets au cours de notre travail.

central et fondamental aux équivalences de la forme

L'énoncé « la neige est blanche » est vrai
si et seulement si
la neige est blanche[3].

Bien souvent, le déflationniste se distingue spécifiquement par le fait qu'il tient non seulement les équivalences formées sur le modèle ci-dessus pour vraies, mais qu'il considère en outre que, prises ensemble, ces équivalences fournissent une analyse exhaustive et complète de la notion de vérité. Elles en fournissent une analyse exhaustive et complète au sens où, à elles seules et sans qu'il soit besoin de fournir d'explication supplémentaire, ces équivalences permettent de rendre compte de tous les emplois légitimes que l'on peut faire de la notion de vérité au sein de nos discours scientifiques ou philosophiques. Cette conception radicalement modeste de la nature de la vérité et cette importance attribuée à l'équivalence entre un énoncé et son attribution de vérité forment le socle commun des déflationnistes. À partir de ce socle, les élaborations et les formulations, tout comme les leçons philosophiques qui en sont tirées, divergent selon les auteurs.

Une première tâche qui nous occupera consistera donc à examiner plus en détails et à clarifier les thèses propres aux auteurs déflationnistes contemporains. C'est à ce travail qu'est consacré notre premier chapitre (1) où, à travers une brève étude historique, nous proposons une tentative de généalogie du déflationnisme. En exposant les idées d'auteurs ouvertement déflationnistes ou qui furent pour eux des sources d'inspiration majeures, on peut percevoir comment certains thèmes touchant la vérité sont apparus dans les réflexions des philosophes puis se sont développés et infléchis pour revêtir la forme qui est la leur au sein du déflationnisme moderne. Nous verrons en particulier que l'une des caractéristiques des déflationnistes actuels est de dénier au prédicat de vérité tout contenu explicatif propre, tout en lui attribuant néanmoins un rôle expressif indispensable qui le rend inéliminable de nos discours théoriques.

Une fois cet effort de clarification effectué, la suite de notre travail se scinde en deux parties d'inégales longueurs. Dans une première partie (2) nous analysons la thèse, souvent attribuée aux déflationnistes contemporains, selon laquelle la vérité serait une

3. Ou bien, si l'on privilégie les propositions comme porteurs de vérité :

La proposition que la neige est blanche est vraie
si et seulement si
la neige est blanche.

sorte de notion logique. Plus précisément, nous examinerons un certain type d'arguments qui ont été fournis à l'appui de cette thèse et nous tenterons de les réfuter. La démarche consiste essentiellement à introduire un critère de logicité permettant de délimiter le champ des notions logiques, puis d'examiner si et dans quelle mesure un prédicat de vérité déflationniste le satisfait. Malgré les tentatives élaborées pour défendre la logicité de la vérité, nous verrons que, selon nous, un prédicat de vérité déflationniste ne satisfait pas les critères (inférentialistes) de logicité.

La seconde partie de notre travail est consacrée à l'évaluation de la portée d'un argument récemment avancé contre les théories déflationnistes. Cet argument, baptisé « argument de la conservativité » ou « argument de la réflexion » consiste à placer le déflationniste face à un dilemme qui paraît insoluble. Selon ses partisans de cet argument, les thèses déflationnistes devraient se traduire par une contrainte de conservativité portant sur leurs théories de la vérité. Mais cette première contrainte apparaît incompatible avec une seconde contrainte, dite d'adéquation, d'après laquelle nos théories de la vérité doivent permettre de formaliser certains raisonnements sémantiques apparus en marge des phénomènes d'incomplétude gödeliens. Dans le chapitre 3, nous exposons précisément l'argument en redonnant chacune de ses étapes et en rappelant les résultats techniques indispensables à sa bonne compréhension. Cet argument a donné naissance à divers débats opposant *pro* et *anti*-déflationnistes. Dans le chapitre suivant (4), nous examinons chacune des voies de réponse qui ont été explorées au nom du déflationnisme pour tenter de sortir du dilemme imposé par l'argument. Nous verrons qu'elles s'articulent selon trois axes principaux correspondant chacun à une prémisse de l'argument. Au terme de cet examen, les conclusions auxquelles nous parvenons, sans prétendre être définitives, ne sont guère favorables au déflationnisme.

On peut considérer ces deux parties de notre travail comme deux tentatives indépendantes mais complémentaires de fournir un cadre méthodologique permettant d'analyser rigoureusement les idées déflationnistes. Par l'emploi de méthodes formelles, il devient possible d'évaluer leurs thèses en s'appuyant sur des résultats techniques précis. Ces deux parties sont de longueurs inégales pour la raison simple que la thèse de la logicité de la vérité, souvent évoquée en passant par les auteurs déflationnistes, n'a que rarement fait l'objet de discussions approfondies et techniquement informées. À l'inverse, l'argument de la conservativité a suscité de très nombreux commentaires et s'appuie sur de multiples résultats de logique mathématiques dont certains sont connus depuis les années 1930.

Notre travail reflète donc cet état de la littérature.

Avant qu'il ne s'engage plus avant dans la lecture de notre travail, il nous faut encore adresser un double avertissement au lecteur : dans ce qui suit, il *ne* sera *pas* question des paradoxes ni du problème des porteurs de vérité, sauf de manière incidente. Il est bien connu que la vérité est impliquée dans de nombreux paradoxes dits sémantiques, dont le plus célèbre est peut-être celui du Menteur [4]. L'étude de ces paradoxes et la question de savoir comment y remédier est devenue centrale pour notre compréhension de la notion de vérité. Elle implique habituellement de quitter le cadre de la logique classique pour examiner les propriétés d'un prédicat de vérité applicable en toute généralité à n'importe quel énoncé, y compris aux énoncés qui contiennent déjà une occurrence du prédicat « vrai ». Ce type de démarche n'est pas sans incidence sur la manière dont on peut comprendre le déflationnisme aléthique [5]. Bien des choses que nous abordons ci-dessous prennent une autre couleur lorsqu'on les approche dans un cadre comprenant un prédicat de vérité « non-typé [6] ». Toutefois, traiter ces questions demanderait un (vaste) travail à part que nous n'avons pas entrepris ici. La compréhension des attributions de vérité (et du déflationnisme) dans les cas les moins problématiques nous semble de toute manière un nécessaire préalable avant d'aborder une réflexion sur les paradoxes. Nous avons donc pris le parti de nous cantonner à un prédicat de vérité « typé », c'est-à-dire ne s'appliquant qu'à des énoncés qui ne contiennent pas déjà eux-mêmes du vocabulaire sémantique.

Un second aspect important des théories philosophiques de la vérité que nous ne traiterons pas ici est la question de la nature des porteurs de vérité. Quelles sont donc ces choses dont nous disons parfois qu'elles sont vraies (ou fausses) ? Les candidats abondent pour tenir ce rôle : propositions, énoncés, énonciations, croyances, états mentaux... la liste est non-exhaustive. Les débats sur ce sujet se sont récemment cristallisés autour de l'opposition entre propositions et énoncés, sans qu'aucun choix ne s'impose. La décision théorique de privilégier tel ou tel porteur de vérité n'est pas sans conséquence pour le déflationniste. Elle est d'ailleurs d'une importance toute particulière pour les programmes déflationnistes inspirés du naturalisme quinien et de son empirisme radical [7].

4. Voyez le chapitre 1 et l'exposé des travaux de Tarski pour une présentation rapide de ce paradoxe page 50.

5. Voyez par exemple Beall et Armour-Garb (2005) pour un recueil de textes récents portant sur cet aspect du déflationnisme

6. c'est-à-dire un prédicat de vérité applicable à des énoncés dans lesquels le prédicat de vérité lui-même est déjà présent.

7. Voyez notre exposition des travaux de Quine 1.2.1 et de Field 1.2.3.

Toutefois, les débats que nous examinons dans notre travail sont largement décorrelés de cette question. Conformément à la pratique des protagonistes de ces discussions, nous avons donc choisi de parler de vérité à propos d'énoncés, sans plus de précaution. Ce que nous disons pourrait néanmoins être facilement reformulé en parlant de vérité des propositions[8] sans que cela change quoi que ce soit à nos arguments. Le seul lieu de notre travail où nous porterons plus particulièrement attention à cette distinction des divers porteurs de vérité sera notre chapitre historique. Dans ce chapitre, nous avons pris soin d'illustrer fidèlement chaque parti adopté par les auteurs dont nous exposons les conceptions, y compris sur cette question. C'est à présent par ce chapitre que nous allons débuter.

8. ou d'autres porteurs de vérité.

Chapitre 1

Déflationnisme et déflationnistes

Au cours de ce chapitre, nous examinons l'évolution des thèmes et des idées philosophiques qui ont donné naissance au déflationnisme moderne en matière de vérité. Pour cela, nous exposons les conceptions de divers auteurs qui ont contribué de manière significative au développement de ces thèses. Sans prétendre à l'exhaustivité, il s'agit donc de proposer ici une brève histoire du déflationnisme à travers ceux qui l'ont faite. Dans cette tentative de généalogie, nous avons délibérément choisi de nous appuyer principalement sur les textes d'origine, que nous avons abondamment cités, de manière à rester le plus proche possible des formulations et des particularités propres à chaque auteur. Par ailleurs, nous avons globalement suivi la chronologie à une importante exception près : nous avons regroupés directement à la suite de Ramsey l'ensemble des auteurs qui se réclament de son héritage pour proposer des théories éliminativistes —ou si l'on préfère des théories de la redondance— de la vérité. Ceci nous amène à aborder les textes élminativistes des années 1970 ou 1990 avant d'exposer la théorie décitationnelle mise au point par Quine à la fin des années soixante. Notre découpage en deux parties ne recouvre donc pas exactement le temps historique. Ce parti-pris s'explique par la logique suivante : selon nous, ce qui distingue les déflationnistes contemporains, c'est l'importance qu'ils accordent au prédicat de vérité comme outil expressif permettant de formuler certaines généralisations. Cette importance explique qu'ils n'imaginent pas que ce prédicat puisse être éliminable ou redondant. Elle les conduit également à vouloir conserver sa forme prédicative à la vérité de manière à permettre son interaction avec la quantification objectuelle. Ce point fondamental distingue les déflationnistes actuels [1],

1. qu'on a parfois qualifiés de « nouvelle vague » déflationniste ; *cf.* Damnjanovic (2010).

sur lesquels porte la suite de notre travail, des partisans de théories éliminativistes de la vérité. C'est la raison pour laquelle nous avons placé l'ensemble de ces derniers parmi les « précurseurs » du déflationnisme moderne, aux côtés de Frege, Ramsey ou Tarski, quitte à faire une entorse à la chronologie.

1.1 Aux origines du déflationnisme

1.1.1 Frege

Ouvrir une histoire du déflationnisme en matière de vérité en y faisant figurer le nom de Gottlob Frege a de quoi surprendre. À première vue, les conceptions de cet auteur touchant la vérité sont en effet diamétralement opposées aux idées qui guident les déflationnistes. Frege fut ainsi l'un des premiers à proposer d'identifier le sens [*Sinn*] d'un énoncé [*Satz*] [2] avec ses conditions de vérité. Il fut également l'un des premiers à traiter les connecteurs propositionnels comme des fonctions (sur les valeurs) de vérité [3]. Mais la thèse la plus célèbre et peut-être la plus intrigante de Frege concernant la vérité est sans doute celle qui consiste à affirmer que la référence [*Bedeutung*] [4] d'une assertion est sa valeur de vérité, à savoir le Vrai ou le Faux [5], et à considérer que le rapport d'une assertion à sa valeur de vérité est en tout point semblable au lien qui unit un nom propre à l'objet qu'il désigne. De ce fait, la notion de vérité joue également un rôle important et particulier dans la théorie frégéenne du jugement puisque Frege tient que « le jugement est non pas la simple saisie d'une pensée mais la reconnaissance de sa valeur de vérité [6] », reconnaissance qui s'effectue non pas en attribuant la propriété de vérité à une pensée mais plutôt en opérant le passage d'un certain niveau, celui des pensées, à un autre niveau, celui des dénotations [7]. Bref, il semble indéniable qu'au delà des emplois prédicatifs du mot « vrai », dans la conception frégéenne de la logique et du langage, la notion de vérité est une notion essentielle amenée à jouer un rôle explicatif fondamental [8].

2. Sens d'un énoncé que Frege appelle également parfois « contenu jugeable » [*beurteilbarer Inhalt*], ou « pensée » [*Gedanke*].

3. Au point que certains, à savoir Heck Jr et May, ont proposé de considérer Frege comme le « grand-père » fondateur de la sémantique (vériconditionnelle) formelle, le père fondateur étant alors Tarski. *Cf.* Heck Jr et May (à paraître, p 1).

4. qu'on traduit aussi parfois par « dénotation ».

5. ces deux valeurs de vérité étant dotées du statut d'*objets* dans l'ontologie frégéenne.

6. Frege (1892, p. 110, note de bas de page, la pagination renvoie à la traduction française).

7. Voyez Frege (1892, p. 110-111).

8. Pour une synthèse récente sur la place de la vérité dans la philosophie de Frege, on pourra consulter Heck Jr et May (à paraître).

C'est donc à titre parfaitement juste que Field (1994a, § 1, p. 104-108) classe Frege parmi les partisans paradigmatiques de ce qu'il appelle la tradition « inflationniste » en matière de vérité [9], tradition que Field *oppose* aux approches déflationnistes dont il est lui-même partisan [10].

Malgré cela, un certain nombre de remarques formulées par Frege peuvent apparaître, *prises isolément*, comme les prémices d'une réflexion déflationniste sur la vérité. C'est pourquoi il n'est pas rare de voir surgir la figure de Frege dans les tentatives de généalogie du déflationnisme [11]. Nous nous conformerons à cet usage et dans les quelques lignes qui suivent nous nous concentrerons simplement sur les passages de Frege qui ont pu être des sources d'inspirations pour les déflationnistes. Ces passages se trouvent dans divers écrits publiés par Frege ainsi que dans certaines de ses œuvres posthumes. En voici un premier exemple tiré de *« Über Sinn und Bedeutung »* publié en 1892 :

> On pourrait être tenté de voir dans le rapport de la pensée au vrai, non pas celui du sens à la dénotation, mais celui du sujet au prédicat. On pourrait dire à cet effet « la pensée que 5 est un nombre premier est vraie ». À regarder la chose de plus près, il apparaît qu'*on a en fait rien dit de plus* que dans la proposition « 5 est un nombre premier ». Dans les deux cas, l'affirmation de la vérité réside dans la forme de la proposition affirmative. Par suite, pour peu que l'affirmation n'ait pas sa force habituelle, par exemple dans la bouche d'un acteur sur scène, la proposition « la pensée que 5 est un nombre premier est vraie » ne contient jamais qu'une pensée, la même que le simple énoncé « 5 est un nombre premier ». Il faut donc admettre que *le rapport de la pensée au vrai ne peut être comparé à celui du sujet au prédicat.* (Frege, 1892, p. 110-111, la pagination renvoie à la traduction française, nous soulignons)

On aperçoit déjà ici deux thèmes qui seront chers aux déflationnistes. Le premier concerne la « transparence » ou la « neutralité » des emplois prédicatifs du mot « vrai ». Puisque, selon les mots de Frege, lorsqu'on ajoute ce qualificatif à la pensée que 5 est un nombre premier, on n'a en fait « rien dit de plus » que dans la proposition « 5 est un nombre pre-

9. Outre Frege, Field range dans cette tradition Russell, le premier Wittgenstein et Ramsey. Cette tradition se caractérise aux yeux Field par le fait qu'elle attribue aux conditions de vérité « un rôle extrêmement central dans la sémantique et dans la théorie de l'esprit » et qu'elle considère qu'« une théorie de la signification ou du contenu est, au moins pour une large part, une théorie des conditions de vérité » (Field, 1994a, p. 104).

10. *Cf. infra* pour une description plus détaillée des positions de Field lui-même.

11. Stoljar et Damnjanovic (2014, § 1) Halbach (2001c) Horwich (1998b) Galinon (2010) en sont quelques exemples parmi d'autres.

mier », il y a, semble-t-il, équivalence —voire identité— de contenu ou de signification, entre la pensée exprimée par un énoncé et celle exprimée par l'énoncé qui lui attribue la vérité. Le second thème introduit l'idée que la vérité ne serait pas une véritable propriété, ou à tout le moins pas une propriété au sens habituel du terme puisque le « rapport de la pensée au vrai ne peut être comparé à celui du sujet au prédicat ». Ces deux thèses, nous allons le voir, seront décisives pour le développement du déflationnisme.

On en retrouve la trace chez Frege lui-même dans un texte plus tardif datant de 1918, tiré des *« Logische Untersuchungen »* et intitulé « La pensée ». Dans ce texte, Frege expose ses vues sur la nature et le rôle de la logique. Après avoir assigné « pour tâche à la logique de trouver les lois de l'être vrai [12] », il entreprend de « dessiner grossièrement les contours de ce qu'[il entendra] par vrai dans la suite de ce texte [13] ». Il parvient tout d'abord, au moyen d'un argument de régression que nous ne détaillerons pas ici, à un résultat d'indéfinissabilité : « toute [...] tentative pour définir l'être vrai échoue [...]. Il est donc vraisemblable que le contenu du mot « vrai » est unique en son genre et indéfinissable [14] ». Puis après avoir précisé qu'il appellera « pensée [*Gedanke*] ce dont on peut demander s'il est vrai ou faux [15] », il aboutit au passage qui nous intéresse plus particulièrement ici :

> Au demeurant, il y a tout lieu de penser que nous ne pouvons pas reconnaître qu'une chose a une certaine propriété, sans en même temps estimer vraie la pensée que cette chose a cette propriété. Ainsi à toute propriété d'une chose est liée une propriété d'une pensée, à savoir celle d'être vraie. Il vaut aussi de remarquer que la proposition « je sens une odeur de violette » a *même contenu* que la proposition « il est vrai que je sens une odeur de violette ». Il semblerait que *rien n'est ajouté* à la pensée quand je lui attribue la propriété d'être vraie. Et pourtant n'est-ce pas un progrès d'importance quand, après une longue hésitation et des recherches pénibles, le savant peut dire enfin « ce que je présumais est vrai » ? La dénotation du mot « vrai » semble *unique en son genre*. Serait-ce que nous ayons affaire à quelque chose qui ne peut nullement être *appelé propriété dans le sens usuel* ? (Frege, (1918-1923), p. 174, nous soulignons)

12. Frege, (1918-1923), p. 171.
13. Frege, (1918-1923), p. 171.
14. Frege, (1918-1923), p. 172-173.
15. Frege, (1918-1923), p. 173.

On retrouve dans cet extrait les deux thèmes précédemment identifiés : celui de l'identité de contenu d'une proposition p et de la proposition « il est vrai que p », ainsi que les doutes soulevés à propos de la vérité qui ne serait pas une propriété au « sens usuel ». Toutefois, Frege lui-même souligne ici qu'il y a un paradoxe apparent entre le fait que prédiquer la vérité n'ajoute rien à nos pensées, alors que cela semble être un progrès notable lorsque « le savant peut dire enfin « ce que je présumais est vrai » ». Pour dissiper ce trouble, Frege développe son analyse et rappelle qu'il convient de distinguer soigneusement la simple expression d'une pensée de son affirmation. Pour illustrer ce point, considérez les trois énoncés suivants :

E_1 : Je sens une odeur de violette.

E_2 : Est-ce que je sens une odeur de violette ?

E_3 : Si je sens une odeur de violette, alors je suis dans le jardin.

Ils contiennent tous trois la pensée que je sens une odeur de violette. Cependant, seul le premier énoncé (E_1), par sa forme déclarative, contient en plus l'affirmation de cette pensée.

Frege considérait comme l'une de ses découvertes fondamentales la mise au jour de cette distinction cruciale entre, d'une part, la simple formulation ou l'expression d'un contenu (contenu qu'on peut retrouver dans une interrogation ou dans une proposition subordonnée, sans qu'il soit affirmé) et, d'autre part, l'assertion ou l'affirmation de ce contenu. Elle est généralement connue aujourd'hui sous le nom de point de Frege depuis que Geach (1965) en a souligné l'importance. Une conséquence de ce point est qu'aucune expression du langage ordinaire ne peut être porteuse par son seul contenu de force assertorique : quel que soit le contenu exprimé par une phrase, celui-ci peut encore être affirmé, supposé, mis en doute, faire l'objet d'une interrogation, placé en antécédent dans une implication, *etc.*

Dans le texte qui nous concerne ici, Frege résume ainsi la situation :

> Les propositions interrogatives et les affirmatives contiennent la même pensée, mais la proposition affirmative contient quelque chose en plus : l'affirmation. [...] Dans une proposition affirmative, il faut donc distinguer deux choses : le contenu qu'elle partage avec l'interrogative correspondante et l'affirmation. [...] On distinguera donc :
>
> 1. La saisie de la pensée — l'acte de penser.

2. La reconnaissance de la vérité d'une pensée — le jugement.

3. La manifestation de ce jugement — l'affirmation.

(Frege, (1918-1923), p. 175)

Dès lors, pour en revenir à notre exemple du savant, Frege poursuit :

> La démarche scientifique comporte d'habitude plusieurs étapes. Il y a d'abord conception d'une pensée [...] ; puis, au terme d'une recherche, on reconnaît que cette pensée est vraie. La reconnaissance de la vérité est enfin exprimée dans la forme de la proposition affirmative. (Frege, (1918-1923), p. 176)

Ainsi, ce qui constitue un progrès pour le savant c'est que, pour fruit de ses recherches, celui-ci est finalement en position non pas seulement de *présumer* que p, mais bien d'en reconnaître la vérité, c'est-à-dire de *juger* que p, et donc de l'*affirmer*.

Mais, et c'est ici que les thèmes déflationnistes affleurent à nouveau, Frege ajoute aussitôt :

> Il n'est nul besoin pour cela de mot « vrai ». Quand bien même l'emploierait-on, la force proprement affirmative ne réside pas en lui mais dans la forme de la proposition affirmative ; si la proposition perd sa force affirmative, le mot « vrai » ne peut pas la lui rendre.
>
> (Frege, (1918-1923), p. 176)

Il est en effet clair qu'à elle seule, la présence du mot « vrai » ne suffit pas à revêtir une proposition de force assertive. Considérons la proposition : « il est vrai que je sens une odeur de violette ». Je puis bien entendu affirmer :

E_1^* Il est vrai que je sens une odeur de violette.

Mais je pourrais tout aussi bien douter qu'il soit vrai que je sens une odeur de violette. Ou encore, demander :

E_2^* Est-il vrai que je sens une odeur de violette ?

Ou bien encore placer cette proposition en antécédent d'une implication pour former un énoncé :

E_3^* S'il est vrai que je sens une odeur de violette alors je suis dans le jardin

dans lequel le contenu de cette proposition, quoique bien présent, n'est ni affirmé ni l'objet d'une question.

Frege conclut donc :

> Que le mot « vrai » soit ou non prononcé, cela n'y change rien. De là, vient que rien ne semble avoir été ajouté à la pensée quand on lui attribue la propriété d'être vraie.

ce qui constitue bien une conclusion aux accents déflationnistes.

D'autres textes de Frege renferment des passages touchant la vérité semblables à ceux que nous venons de citer, notamment parmi ses écrits posthumes. D'ailleurs, c'est peut-être dans un court texte non publié du vivant de l'auteur et dont la date de rédaction est estimée à 1915 que l'on trouve les observations de Frege les plus radicales et les plus proches d'une théorie éliminative de la vérité en bonne et due forme. Dans ce texte intitulé *Mes intuitions logiques fondamentales*, Frege entend proposer « une clef donnant accès à l'intelligence des résultats auxquels [il est] parvenu [16] ». Ses analyses sur la vérité tiennent une place centrale dans cette entreprise et Frege déclare :

> Le mot « vrai » n'apporte donc par son sens aucune contribution essentielle à la pensée. Quand j'asserte « il est vrai que l'eau de mer est salée », j'asserte la même chose que quand j'asserte « l'eau de mer est salée ».

Revoici le thème de la trivialité ou de la transparence des attribution de vérité. Mais cette fois Frege l'accompagne de remarques franchement négatives concernant la nature de la vérité. Tout d'abord :

> On pourrait dès lors être d'avis que le mot « vrai » ne possède simplement aucun sens. Mais la phrase dans laquelle le mot « vrai » est prédicat ne possèderait alors non plus aucun sens. On peut simplement dire : le mot « vrai » a un sens qui n'apporte rien au sens de la phrase entière dont il est le prédicat. (Frege, 1915, p. 298)

Et quelque lignes plus loin :

> [...] le mot « vrai » ne fait en réalité qu'une tentative malheureuse pour indiquer [la nature] de la logique, dans la mesure où ce qui est réellement en question ne se rapporte pas du tout au mot « vrai » mais à la force assertive avec laquelle une phrase est prononcée. (Frege, 1915, p. 298)

Enfin, dans les ultimes remarques de son texte, un pas supplémentaire est franchi puisqu'ici Frege semble bel et bien plaider pour l'élimination pure et simple des emplois de la notion de vérité :

16. Frege, 1915, p. 297.

> D'où provient dès lors le fait que le mot « vrai », quoiqu'il semble vide de contenu, soit néanmoins indispensable ? S'agissant de fonder la logique ne pourrait-on ici au moins l'éviter entièrement, puisqu'il peut seulement engendrer la confusion ? Le fait que cela ne nous soit pas possible provient de l'imperfection de la langue. Si nous avions une langue logiquement parfaite, nous n'aurions peut-être plus aucun besoin de la logique ou alors nous pourrions la lire dans la langue elle-même. (Frege, 1915, p. 298)

Ainsi, les inévitables (et trompeurs) emplois du prédicat de vérité ne proviendraient que des imperfections des langues naturelles, imperfections auxquelles il serait souhaitable de remédier. À ce stade on ne peut qu'être frappé de la proximité des analyses de Frege avec celle qui seront développées quelques années plus tard par Ramsey et qui seront reprises par les théoriciens de la vérité-redondance.

En résumé, si vouloir faire de Frege un déflationniste serait sans doute commettre à la fois un anachronisme et un contresens, on peut toutefois relever chez lui bien des remarques dont le contenu anticipe des thèmes qui seront développés ultérieurement par les déflationnistes. Plus précisément, on découvre déjà sous sa plume la thèse de la transparence des attributions du prédicat « vrai », les premiers doutes concernant le statut d'authentique propriété de la notion de vérité et même ce qu'on peut qualifier de « première forme connue de théorie de la vérité redondance [17] ». Pour poursuivre notre exposé, nous allons à présent tourner notre regard vers un auteur qui, s'il n'était pas non plus lui-même déflationniste, a également exercé une influence déterminante sur les auteurs déflationnistes qui l'ont suivi. Il s'agit de Frank Ramsey.

1.1.2 Ramsey et les prophrases

Frank Ramsey a souvent été considéré comme un précurseur et un représentant typique de la théorie de la vérité-redondance, et, à ce titre, comme un auteur déflationniste avant l'heure. Pourtant cette classification est sans doute erronée. Comme l'ont montré Field (1986) et Rivenc (1998), les conceptions de Ramsey concernant la vérité le rangeraient plutôt du côté des théoriciens de la correspondance. Il faut signaler que l'interprétation des écrits de Ramsey sur la vérité s'est avérée d'autant plus difficile qu'une partie importante de ses réflexions sur le sujet n'a été divulguée que très tar-

17. Selon une formulation que nous empruntons à l'avant-propos des directeurs de la traduction des *Écrits posthumes* de Frege. *Cf.* Frege (1999, p. I).

divement et à titre posthume. Si les remarques sur la vérité contenues dans *Facts and propositions* (Ramsey, 1927) ont été publiées du vivant de l'auteur, le manuscrit *On truth* (Ramsey, 1991), sans doute composé vers la même époque, a dû attendre plus de soixante ans pour être enfin révélé au public. Pourtant, c'est seulement à la lumière de ces deux textes qu'on peut pleinement saisir le propos de Ramsey sur la vérité dans la mesure où ce second manuscrit prolonge et approfondit les observations du précédent. [18] Avant d'en venir aux aspects de sa théorie qui ont pu être interprétés comme fournissant les éléments d'une théorie déflationniste et qui sont ceux qui nous intéressent principalement ici, nous commençons par expliquer pourquoi Ramsey était sans doute en réalité un théoricien de la correspondance.

Dans Ramsey (1927), Ramsey propose une analyse du jugement dans laquelle il distingue deux types de facteurs : il y a d'une part ce qu'il appelle les facteurs objectifs qui sont les faits, les propositions ou les choses du monde, et d'autre part ce qu'il baptise les facteurs mentaux et qui sont « mon esprit, ou mon état mental présent, ou encore les mots ou les images dans mon esprit » [19]. Selon lui,

> il est naturel de supposer que le fait que je suis en train de juger que César fut assassiné consiste en l'existence d'une certaine relation, ou de certaines relations, entre ces facteurs mentaux et ces facteurs objectifs. (Ramsey, 1927, p. 215)

Le problème que se pose Ramsey ici consiste donc à analyser le genre de relation qui relie les deux types de facteurs dans un jugement : il s'agit d'expliquer quand et comment des facteurs objectifs, disons le fait que César a été assassiné, sont exprimés ou représentés par des facteurs mentaux tels que la croyance, le jugement ou l'assertion que César a été assassiné. Selon Ramsey l'explication de cette relation de représentation doit se faire en termes de correspondance.

L'analyse du concept de vérité proprement dit est, pour Ramsey, inséparable de cette explication de la relation de représentation entre facteurs mentaux et facteurs objectifs. Selon Ramsey (1927) en effet la notion de vérité n'est pas (ou pas directement) attribuée aux facteurs mentaux tels que la croyance. Elle s'applique plutôt aux facteurs objectifs, à savoir la proposition [20] exprimée ou représentée par le facteur mental. En d'autres

18. Ajoutons qu'il semble malheureusement aussi parfois le contredire, ce qui ne facilite pas l'interprétation.
19. Ramsey, 1927, p. 215.
20. Ou peut-être faudrait-il dire à la *référence propositionnelle*, si l'on en croit Ramsey (1991) (nous

termes, toujours selon Ramsey, une croyance est vraie si la proposition qu'elle exprime est vraie. Pour mener à bien l'analyse du concept de vérité, Ramsey se trouve donc face à deux tâches : d'une part, expliquer en quoi consiste pour une croyance (ou plus généralement pour un facteur mental) d'être une croyance en telle ou telle proposition, et d'autre part, dire ce que cela signifie pour une proposition qu'être vraie. La réponse apportées par Ramsey à la première de ces deux questions en fait un théoricien de la correspondance, tandis que celle qu'il apporte à la seconde et qui culminera sous la forme d'une définition de la vérité, contient les ferments d'une analyse déflationniste de la vérité.[21] En gardant à l'esprit cette articulation en deux temps de l'analyse de Ramsey, nous sommes en mesure de l'examiner plus en détails.

Une première difficulté rencontrée est celle de la nature des porteurs de vérité. Dans *Facts and propositions* (1927), Ramsey attribuait la vérité aux propositions :

> La vérité et la fausseté se disent avant tout de propositions. (Ramsey, 1927, p. 217) :

Dans Ramsey (1991), il revient sur cette question et y apporte un nouvel éclairage. Dorénavant Ramsey considère comme problématique et philosophiquement douteuse l'existence de propositions « objectives ». Pour autant, il n'entend pas appliquer la vérité directement aux énoncés ou aux assertions puisqu'il est clair à ses yeux que la vérité d'un énoncé dépend de sa signification. Il propose donc de considérer certains « états mentaux »[22] comme porteurs de vérité :

---

allons y revenir). La question de la nature des porteurs de vérité est manifestement l'un des points sur lesquels Ramsey a varié entre (Ramsey, 1927) et (Ramsey, 1991).

21. C'est peut-être à Field qu'il revient d'avoir résumé de la manière la plus concise ce double aspect des analyses de Ramsey :

> Une conception qui est clairement *non* déflationniste est celle de Ramsey : Ramsey est un cas clair de théoricien de la correspondance. Bien sûr, Ramsey ne pensait pas que *le mot « vrai »* exprime une notion de correspondance : il considérait que ce mot s'applique à ce qu'il appelait des « propositions », qui sont simplement des encapsulations des conditions de vérité ; et que ainsi appliqué, il est redondant, au sens où dire de la proposition que César a franchi le Rubicon qu'elle est vraie, c'est simplement dire que César a franchi le Rubicon. Mais Ramsey reconnaissait que cela ne rendait pas trivial le problème de la vérité [...] : il concluait que « le problème n'est pas celui de la nature de la vérité ou de la fausseté mais celui de la nature du jugement ou de l'assertion... » (Ramsey, 1927). C'est-à-dire que le problème réel concerne ce en quoi consiste pour une affirmation ou un état de pensée d'*exprimer la proposition que* César a franchi le Rubicon, c'est-à-dire d'*avoir pour conditions de vérité que* César a franchi le Rubicon. (Field, 1986, p. 60, italiques de l'auteur)

22. qui, à première vue, rappellent évidemment les facteurs mentaux de Ramsey (1927).

> Notre tâche, donc, est de clarifier les termes vrais et faux [23] en tant qu'ils s'appliquent à des états mentaux, et comme exemple typique de ces états qui nous occupent nous pouvons prendre pour le moment les croyances. Dès lors, qu'il soit ou non philosophiquement correct de dire qu'elles ont pour objet des propositions, les croyances ont sans aucun doute une caractéristique que je me risquerai à nommer *référence propositionnelle*. Une croyance est nécessairement une croyance que telle chose ou telle autre est telle ou telle, par exemple que la terre est plate ; et c'est cet aspect, celui d'être une croyance « que la terre est plate » que je propose d'appeler sa référence propositionnelle. (Ramsey, 1991, p. 7, italiques de l'auteur) [24]

Toutefois, le simple fait pour un état mental de posséder une référence propositionnelle est insuffisant pour pouvoir se voir attribuer une valeur de vérité. Encore faut-il que cet état mental possède une dimension affirmative :

> D'un autre côté, tous les états mentaux possédant une référence propostionnelle ne sont pas soit vrais, soit faux. [...] Nous n'appelons pas vrais des souhaits, des désirs ou des interrogations, non parce qu'ils n'ont pas de référence propositionnelle, mais parce qu'il leur manque ce qu'on pourrait appeler un caractère affirmatif ou assertif, cet élément qui est présent dans le fait de penser que, mais absent lorsqu'on se demande si. (Ramsey, 1991, p. 8)

Les porteurs de vérité seront donc des états mentaux possédant, outre une référence propositionnelle, ce caractère affirmatif. Ramsey propose un emploi technique des termes « croyance » et « jugement » pour les désigner :

> Les états mentaux qui nous concernent, à savoir, précisément, ceux ayant une référence propositionnelle et un certain degré de caractère affirmatif, n'ont malheureusement pas de nom commun dans le langage ordinaire. Il n'existe pas de terme applicable à l'ensemble du spectre allant de la simple conjecture à la connaissance certaine, et je propose de pallier ce manque en employant les termes de *croyance* et de *jugement* de manière synonyme pour couvrir l'ensemble du spectre des états mentaux en question, bien que cela

23. N.D.T. sans guillemets dans le texte original.
24. Nos traductions de Ramsey (1991) sont pour partie reprises de Rivenc (1998) qui en cite de nombreux extraits.

> nécessite une grande extension de leur signification ordinaire, plutôt que dans leur sens ordinaire plus restreint.
>
> C'est donc au regard des croyances ou des jugements que nous nous demandons ce que signifient la vérité et la fausseté [...] (Ramsey, 1991, p. 8, italiques de l'auteur)

À la lumière des extraits ci-dessus, on constate donc une évolution notable dans les réflexions de Ramsey entre l'article de 1927 et le manuscrit posthume publié en 1991 : dans un premier temps Ramsey attribuait la vérité aux propositions, qu'il rangeait parmi les *facteurs objectifs*, alors que dans un second temps il recommande plutôt de l'attribuer aux croyances —en un sens technique de ce terme qu'il a préalablement introduit— c'est-à-dire à des *états mentaux* dotés d'une référence propositionnelle.

Malgré cette importante différence de formulation, il nous semble que la conception de la vérité de Ramsey n'a pas véritablement varié de manière essentielle. La structure générale de son analyse demeure inchangée. On peut proposer l'interprétation suivante qui, mettant en parallèle la formulation en termes de proposition (Ramsey, 1927) et celle en termes de croyance+référence propositionnelle (Ramsey, 1991), permet de donner une vision cohérente commune aux deux textes sur la question de la vérité. Dans les deux cas, les réflexions de Ramsey procèdent en deux étapes ; il sépare

1. premièrement, la question de savoir comment [un facteur mental]/{une croyance} [25] peut être dit [représenter telle ou telle proposition p]/{avoir telle ou telle référence propositionnelle p}, et
2. secondement, la question de savoir ce que signifie attribuer la vérité *une fois donnée* [la proposition]/{la référence propostionnelle}.

Malgré les différences de formulation, sans doute dues principalement aux soupçons de Ramsey quant à l'emploi des propositions « objectives », la stratégie d'analyse reste donc globalement la même. L'introduction de la notion de référence propositionnelle est vraisemblablement une manière simplement prudente d'éviter l'hypothèse ou l'hypostase [26] des propositions « objectives ». Et, comme nous l'avons déjà dit, c'est dans l'examen de la première question que se loge la nature correspondantiste de la position de Ramsey, tandis que ce sont les éléments qu'il développe en réponse à la seconde question qui ont un parfum « éliminativiste » ou « redondantiste » et qui ont pu le faire prendre, sans

25. selon les formulations de [Ramsey (1927)]/ {Ramsey (1991)} respectivement.
26. Selon une formule empruntée à Rivenc (1998, p. 18).

doute à tort, pour un déflationniste. Mais laissons de côté les hésitations de Ramsey sur la nature des porteurs de vérité et concentrons-nous un peu plus sur ce qu'il a à dire à propos des attribution du prédicat « vrai » proprement dites, c'est-à-dire sur la deuxième question ci-dessus.

Dans un passage célèbre de Ramsey (1927, p. 217), Ramsey distingue deux cas pour approcher ce problème : les cas où la [proposition] [27] à laquelle la vérité est attribuée nous est donnée explicitement et les cas où elle fait l'objet d'une description.

Dans le premier cas, Ramsey élabore une forme de théorie éliminative du prédicat de vérité ou théorie de la redondance. Il écrit :

> Supposons d'abord qu'elle soit explicitement donnée ; alors il est évident que « Il est vrai que César fut assassiné » ne signifie rien de plus que César fut assassiné, [...] Ce sont des expressions que nous utilisons quelquefois pour mettre l'accent soit pour des raisons stylistiques, soit pour indiquer la position occupée par l'énoncé dans notre argumentation. (Ramsey, 1927, p. 217)

Dans ce premier type de situation, le prédicat « vrai » est donc redondant au sens où il est éliminable sans perte de signification et où, par conséquent, ses capacités expressives sont en réalité superflues.

Le second cas, celui où les propositions auxquelles la vérité est attribuée ne sont pas données explicitement mais font l'objet d'une description ou d'une quantification, est plus problématique. En effet, une élimination directe par reformulation n'est alors pas possible dans le langage ordinaire. Pour illustrer cette difficulté, Ramsey prend pour exemple une assertion de la forme « Jean a toujours raison » qu'il interprète comme signifiant que les propositions que Jean affirme sont toujours vraies. Contrairement au cas précédent, cette dernière assertion semble impossible à exprimer dans le langage ordinaire sans recourir au prédicat de vérité. Mais, poursuit Ramsey,

> supposez que nous le formulions ainsi : « Pour tout p, s'il affirme que p, alors p est vrai », alors nous voyons que la fonction propositionnelle p est vraie est tout simplement la même chose que p, comme par exemple sa valeur « César fut assassiné est vrai » est la même chose que « César fut assassiné ». Nous avons en français à ajouter « est vrai » pour donner à la phrase un verbe, en oubliant que « p » contient déjà un verbe (variable). (Ramsey, 1927, p. 217)

27. Respectivement : { les cas où la *référence propositionnelle* de la *croyance* (au sens technique) à laquelle la vérité est attribuée... }, pour le dire dans une formulation *à la* Ramsey (1991) (si on veut bien nous permettre d'imaginer ce qu'elle pourrait donner).

Là encore, même dans les cas contenant une quantification sur une classe de propositions, c'est l'équivalence entre la fonction propositionnelle « p est vraie » et la proposition « p » qui est au cœur de l'explication. Ce que suggère la dernière phrase de l'extrait cité, c'est que dans les cas de ce type, la présence et l'inéliminabilité du prédicat de vérité est simplement due au fait que, dans l'énoncé « Pour tout p, s'il affirme que p, alors p est vrai », nous avons toujours une lecture objectuelle de la quantification. Les variables p sont censées désigner des objets et ne peuvent donc pas apparaître en position d'énoncé. Elles sont semblables aux pronoms dans le langage courant. Et cette lecture nous empêche de voir que p, dans la mesure où elle désigne une proposition, contient déjà en quelque sorte un verbe. Pour obtenir une phrase grammaticalement correcte il nous faut donc accoler la copule « est vraie » à la variable pro-nominale p. Mais cette nécessité syntaxique ou grammaticale de la présence du prédicat de vérité ne s'accompagne d'aucune modification ou ajout de contenu. Elle est simplement le produit de la pauvreté expressive du langage ordinaire et de la conception étroitement objectuelle de la quantification usuelle. Immédiatement après le passage cité ci-dessus, Ramsey poursuit d'ailleurs comme ceci, essayant de contourner ce problème :

> Cela sera peut-être plus clair si l'on suppose un moment que seulement une forme de proposition est en cause, en l'occurrence la forme relationnelle aRb ; alors « Il a toujours raison » pourrait être exprimé par « Pour tout a, R, b s'il affirme aRb, alors aRb », à laquelle « est vrai » serait une addition superflue. Quand toutes les formes de propositions sont inclues, l'analyse est plus compliquée mais elle n'est pas essentiellement différente [...] (Ramsey, 1927, p. 217) [28]

Un pas supplémentaire est franchi dans Ramsey (1991). Dans cet autre texte, même s'il ne la développe pas de manière explicite ou entièrement formalisée, Ramsey penche encore plus nettement pour une lecture hétérodoxe de la quantification permettant la quantification sur des énoncés en position d'usage. Et cette démarche lui permet même d'aboutir à une définition de la vérité. Plus précisément, il écrit :

28. Notons qu'en toute rigueur, la formulation de Ramsey dans ce dernier extrait n'est d'ailleurs pas tout à fait correcte. Même si l'on accepte la quantification sur les « objets » a, R, b (qui visiblement contiennent néanmoins une relation), la première occurrence de aRb dans l'énoncé « Pour tout a, R, b s'il affirme aRb, alors aRb » semble être un nom de la proposition affirmée, tandis que la seconde est en position d'usage, au sens où aRb apparaît ici comme une proposition assertée, ou plus précisément comme une proposition reliée à d'autres par un connecteur logique pour former un énoncé plus complexe, mais en tout cas *pas* comme un nom. Il y a donc une confusion ou une ambiguité quant à la nature de aRb : nom d'une proposition ou proposition elle-même ?

> [...] nous devons considérer la forme générale d'une référence propositionnelle [...] ; nous pouvons symboliser toute croyance comme une croyance que p, où « p » est un énoncé variable comme « A » et « B » sont des mots ou expressions variables (ou des termes comme on dit en logique). Nous pouvons alors dire qu'une croyance est vraie si c'est une croyance que p, et p. Cette définition sonne d'abord bizarrement parce que nous ne comprenons pas tout de suite que « p » est un *énoncé* variable et doit être regardé comme contenant un verbe. (Ramsey, 1991, p. 9)

L'analyse du problème consistant à « définir la vérité au sens d'en expliquer la signification » [29] aboutit donc chez Ramsey à une définition explicite de la forme :

(Def) : Une croyance est vraie *ssi* c'est une croyance que p, et p

Cette définition est rendue possible par l'emploi non pas simplement de variables d'énoncés au sens de variables dont les valeurs possibles seraient des énoncés, mais de variables *en position d'énoncé* : dans la définition ci-dessus la seconde occurrence de p prend en effet la place d'un énoncé et non pas d'un nom. Elle « désigne » donc une proposition en position d'usage. Ces variables d'un type nouveau, Ramsey lui-même les qualifie de « pro-phrases » [30].

Ces pro-phrases, ou ces variables pro-phrastiques, sont supposées avoir un rôle analogue au rôle pronominal des variables ordinaires mais pour une autre catégorie d'expression, celle des phrases. Aux yeux de Ramsey ces variables occupant des positions d'énoncés ne posent pas de difficulté particulière. Et il ne donne d'ailleurs aucun détail sur la manière dont il faudrait les formaliser ou sur la sémantique qu'il faudrait leur associer. Si elles nous font sortir du cadre de la logique standard et si la définition ci-dessus n'est pas paraphrasable dans le langage ordinaire, il n'y a là pour Ramsey qu'un problème factice qui résulte de certaines lacunes expressives, à savoir la pauvreté du langage ordinaire en prophrases :

> Dans la mesure où nous prétendons avoir défini la vérité, nous devrions être capables de substituer notre définition au mot « vrai » partout où il figure. Mais la difficulté que nous avons mentionnée rend cela impossible dans le langage ordinaire, qui traite ce qui devrait être en réalité appelé des *pro-phrases* comme s'il s'agissait de *pro-noms*. [...] cette particularité du langage

29. Ramsey, 1991, p. 13.
30. Ramsey, 1991, p. 10 (voyez l'extrait cité ci-dessous).

> donne naissances à des problèmes artificiels quant à la nature de la vérité, problèmes qui disparaissent dès qu'ils sont exprimés dans le symbolisme logique, où nous pouvons rendre « ce qu'il croit est vrai » par « si p est ce qu'il croit, alors p ». (Ramsey, 1991, p. 10)

Ainsi, même dans les cas d'expressions à caractère de généralité où l'on est confronté, au moins apparemment, à une quantification sur une classe infinie de propositions, le prédicat « vrai » est finalement éliminable, au prix de quelques difficultés techniques dans la schématisation quantificationnelle des énoncés du langage naturel. Si l'on accepte de s'éloigner de la grammaire de surface des énoncés étudiés, et si l'on adopte les variables pro-phrastiques, le prédicat de vérité devient explicitement définissable et est donc en fait éliminable et redondant. On n'y aura alors recours que pour des raisons stylistiques et rhétoriques ou pour conserver une uniformité grammaticale apparente.

On voit bien ici comment alors même que, nous l'avons rappelé, Ramsey était vraisemblablement un représentant de la théorie de la correspondance [31] plutôt qu'un véritable déflationniste, la partie éliminative de son analyse du prédicat de vérité a pu être source d'inspiration pour des auteurs éliminativistes en un sens plus fort. Elle a ouvert la voie à des auteurs qui non seulement remettent en cause la syntaxe de surface qui fait

31. Pour être exhaustif sur ce point, signalons que Ramsey est plus explicite sur cet aspect de son travail dans Ramsey (1991) que dans Ramsey (1927). Il écrit par exemple,

> Bien que nous n'ayons pas encore utilisé le mot « correspondance » notre théorie sera probablement appelée une théorie de la correspondance. (Ramsey, 1991, p. 11)

Et la raison en est que la définition explicite de la vérité proposée, celle qui permet l'élimination ou la paraphrase des contextes contenant des prédications de vérité est une

> [...] définition donnée en termes de référence propositionnelle, que nous prenons pour un terme déjà compris. Mais il est permis penser que cette notion de référence propositionnelle a elle-même besoin d'être analysée et définie, et qu'une définition de la vérité dans les termes d'une notion si obscure ne représente qu'un maigre progrès [...] La vérité, dira-t-on, consiste en une relation entre des idées et la réalité, et l'utilisation sans analyse du terme de référence propositionnelle dissimule et esquive simplement les problèmes réels que soulève cette relation. [...]
>
> Nous devons admettre que cette charge est juste, et une explication de la vérité qui accepte la notion de référence propositionnelle sans analyse ne saurait être considérée comme complète. Car toutes les nombreuses difficultés liées à cette notion sont véritablement impliquées dans la vérité qui dépend d'elle ; si par exemple « référence propositionnelle » a une signification très différente selon les différents genres de croyances (comme beaucoup le pensent), alors une ambiguïté similaire est également latente dans « la vérité », et il est évident que nous n'aurons pas clarifié notre idée de la vérité tant que ce problème et tous les problèmes semblables ne seront pas résolus. (Ramsey, 1991, p. 13-14)

On retrouve bien ici l'articulation en deux temps de l'analyse de Ramsey telle que nous l'avons exposée page 18.

de « vrai » une expression prédicative, mais encore récusent l'idée selon laquelle la vérité serait une véritable propriété attribuée à certains porteurs. On perçoit bien en effet chez Ramsey les balbutiements d'une théorie de la redondance du prédicat de vérité articulée autour d'une quantification substitutionnelle. Anticipant quelque peu sur la chronologie, disons quelques mots sur les développements ultérieurs auxquels les réflexions de Ramsey ont donné naissance.

1.1.3 L'héritage éliminativiste de Ramsey

Ayer (1946) fut l'un des premiers auteurs à reprendre et développer les analyses de Ramsey tout en les radicalisant. Selon lui, les explications traditionnelles de la vérité comme « qualité ou relation réelle » sont tout simplement erronées et dues à une mauvaise compréhension des énoncés de notre langage. D'un point de vue ontologique, la vérité n'est tout bonnement pas du tout une propriété, et toute l'« analyse » de ce concept se réduit en fait à un examen correct des énoncés du type « p est vrai » où « vrai » s'applique (ou plutôt semble s'appliquer puisqu'il est en réalité éliminable) à une proposition p. Dans les cas où la proposition est expressément donnée, « vrai » est simplement redondant :

> Revenant à l'analyse de la vérité, nous trouvons que dans toutes les phrases de la forme « p est vrai », l'expression « est vrai » est logiquement superflue. [...] Ainsi, dire qu'une proposition est vraie c'est exactement l'affirmer et dire qu'elle est fausse c'est affirmer sa contradictoire. Cela indique que les termes « vrai » et « faux » ne connotent rien, mais fonctionnent dans la phrase simplement comme signes d'assertion ou de négation. Et dans ce cas, il ne peut y avoir de sens à nous demander d'analyser le concept de « vrai ». (Ayer, 1946, p. 122) [32]

Dans le cas des énoncés généraux, ou universellement quantifiés, on aura recours à une réduction plus élaborée qui s'appuiera sur la totalité des instances de l'énoncé universel dans chacune desquelles « vrai » sera à nouveau directement éliminable. Voici une illustration typique de cette démarche due à Ayer lui-même :

> [...] la phrase : « La vérité est quelquefois plus paradoxale (*stranger*) que la fiction » est équivalente à : « Il y a des valeurs de p et de q telles que p est vrai et q est faux, et que p est plus surprenant que q. » [...] Dans chaque cas, l'analyse de la phrase confirmerait notre hypothèse selon laquelle

32. La pagination renvoie à la traduction française.

> la question : « Qu'est-ce que la vérité ? » est réductible à la question : « Quelle est l'analyse de la phrase : p est vrai ? » Et il est clair que cette question ne soulève aucun problème réel, puisque nous avons montré que dire que p est vrai est simplement une manière d'affirmer p. (Ayer, 1946, p. 123)

Cette construction qui semble pointer vers la quantification substitutionnelle est quasiment semblable à celle de Ramsey[33] auquel Ayer fait d'ailleurs explicitement référence[34]. Ayer est cependant plus radical que Ramsey. Nous seulement, il ignore entièrement la dimension correspondantiste du Ramsey historique[35], mais en outre, d'une analyse grammaticale du prédicat de vérité (« vrai » est redondant et éliminable au prix de quelques détours par la quantification substitutionnelle), Ayer tire une véritable position métaphysique : la vérité n'est pas réellement une propriété, ce n'est q'une « façon de parler », un trait contingent (et trompeur) de nos pratiques linguistiques :

> Nous concluons alors qu'il n'y a pas de problème de la vérité tel qu'il est ordinairement conçu. La conception traditionnelle de la vérité considérée comme une « qualité réelle » ou une « relation réelle » est due, comme la plupart des erreurs philosophiques, à une incapacité d'analyser les phrases correctement. (Ayer, 1946, p. 124)

Ce passage d'une analyse logico-grammaticale de nos usages du prédicat de vérité et de la remise en cause de son caractère véritablement prédicatif à des affirmations concernant la nature ou l'absence de nature de la notion de vérité est un trait commun à de nombreux auteurs déflationnistes ou éliminativistes.

Un autre exemple de remise en cause du statut prédicatif de la notion de vérité nous est fourni par Williams (1976). Cet auteur se rapproche d'Ayer en ce qu'il considère que l'on n'attribue jamais une propriété à quoi que ce soit par l'emploi du mot « vrai ». Selon lui, les usages fondamentaux du prédicat de vérité ne sont pas ceux où « vrai » est prédiqué d'un nom ou d'un objet syntaxique explicitement donné. Ainsi, des énoncés tels que « « La neige est blanche » est vrai » ou bien encore tels que « La proposition que la neige est blanche est vraie » sont à ses yeux peu usuels. Au contraire, les constructions typiques impliquant le prédicat de vérité sont celles où celui-ci apparaît dans des énoncés généraux, des expressions contenant des quantifications, comme par exemple

33. Ou plus précisément à la partie éliminative de la théorie de Ramsey, celle qui correspond à la seconde des deux étapes que nous avons isolées page 18.
34. *cf.* Ayer (1946, p. 123 note de bas de page)
35. Ce que, page 18, nous avons appelé le premier temps de l'analyse de Ramsey.

Ce que dit Percy est vrai.[36]

De ces expressions, on peut donner une analyse éliminative quant au prédicat de vérité, en utilisant, là encore, une forme de quantification qui lorgne vers la quantification substitutionnelle. Ainsi, Williams analyse l'exemple ci-dessus de la manière suivante :

> Pour un certain p, pour tout q, la proposition que p est la même proposition que q si et seulement si Percy dit que q et p. (Williams, 1976, p. 38)

Cette formulation pour le moins maladroite a simplement pour but de s'assurer que Percy a bien dit une seule et unique chose[37]. Si on laisse de côté l'exigence d'unicité, l'analyse de Williams revient à ceci :

Pour un certain p, Percy dit que p, et p.[38]

Dans cette réduction, l'expression prédicative « est vrai » a tout bonnement disparu et a été remplacée par une forme de quantification sur les propositions. C'est là la raison principale qui conduit Williams a rejeter l'idée selon laquelle « est vrai » attribue quelque propriété que ce soit à quelque objet que ce soit. Pour Williams donc, le prédicat de vérité joue plus un rôle de « quantificateur » pour le langage naturel que d' « attributeur » de propriété.

Parmi les courants déflationnistes héritiers de Ramsey, il nous faut enfin mentionner la théorie « prophrastique » de la vérité[39]. Cette théorie a été formulée par Grover (1973, 1972) et Grover, Camp et Belnap (1975)[40]. Il s'agit ici aussi d'une analyse de la vérité en terme de quantification propositionnelle. Elle se présente ouvertement comme une élaboration des idées de Ramsey[41] et donne un rôle central à la notion de

36. Ceci est l'exemple favori de Williams (1976).

37. Selon Halbach (2001c, p. 67) sur lequel nous nous appuyons ici, Williams emploie en fait une construction similaire à celle employée par Russell pour l'analyse des descriptions définies.

38. C'est ce que Williams (1976, p. 1) lui-même dit :

> Dire que ce que Percy dit est vrai, c'est dire que les choses sont comme Percy dit qu'elles sont, *i.e.* (du moins en première approximation) que
>
> Pour un certain p, à la fois Percy dit que p et p.

La parenté avec Ramsey est évidemment frappante ici.

39. Pour une étude critique de la théorie prophrastique de la vérité et une comparaison avec les idées de Ramsey, voyez Rivenc (1998).

40. Grover (1992) est un recueil rassemblant les articles de l'auteur sur ce sujet qui offre une vue globale de la conception prophrastique de la vérité.

41. Voyez Grover, Camp et Belnap (1975, § 1) pour une référence explicite et revendiquée à cet auteur, ainsi que Rivenc (1998) pour une comparaison des approches prophrastiques avec celle de Ramsey.

pro-phrase [42]. Toutefois, contrairement aux remarques souvent elliptiques de Ramsey, ces auteurs s'efforcent d'en proposer une analyse plus longuement développée et détaillée.

À la manière de Ramsey donc, Grover, Camp et Belnap (1975, p. 75) convoquent, pour paraphraser certaines prédications de vérité, une quantification propositionnelle dans laquelle les variables propositionnelles sont susceptibles d'occuper des positions d'énoncés. Ainsi,

(1) Tout ce que dit Jean est vrai

se rend dans un langage semi-formel par

(1') $\forall p$ (Jean dit que $p \rightarrow p$) [43]

Comme nous l'avons déjà signalé à propos de Ramsey, l'énoncé (1') ci-dessus apparaît tout d'abord ***a***grammatical dans une lecture standard de la quantification où les variables sont censées ne pouvoir figurer que dans des positions nominales. Grover, Camp et Belnap (1975) eux-mêmes le reconnaissent :

> Supposez que nous essayons de rendre en anglais simple (1'), qui est formulé dans le langage de Ramsey muni de ses quantificateurs propositionnels. Si nous suivons le style de lecture habituellement attribuée aux formules contenant des occurrences liées de variables individuelles, nous emploierons des pronoms pour saisir la variable liée, ce qui donne
>
> (1") Pour toute proposition, si Jean dit qu'elle, alors elle.
>
> (Grover, Camp et Belnap, 1975, p. 81)

Il est clair que ce dernier énoncé (1") est mal formé. Mais en fait, ce que propose la théorie prophrastique, c'est d'introduire une nouvelle sémantique des quantificateurs [44] : selon Grover, Camp et Belnap —comme Ramsey (1991) l'avait déjà pressenti— , l'apparente *a*grammaticalité de (1') et l'impossibilité supposée à le paraphraser en langue naturelle, sont dues à la nature « prophrastique » des variables de proposition :

42. Sur ce point, sans doute serait-il d'ailleurs plus correct de parler de redécouverte plutôt que d'élaboration des idées de Ramsey puisqu'à l'époque des travaux de Grover, Camp et Belnap, les écrits de Ramsey (Ramsey, 1991) dans lesquels il anticipait et la notion et la terminologie étaient encore restés inédits.

43. *cf.* Grover, Camp et Belnap (1975, p. 75).

44. En anticipant sur ce qui va suivre on peut déjà dire que cette sémantique s'appuie (ou semble s'appuyer) sur une interprétation substitutionnelle des quantificateurs. Voyez ci-dessous.

> Nous avons besoin de quelque- chose qui accomplit le genre de renvoi de référence effectué par les variables, et qui puisse occuper une position qu'une phrase pourrait occuper. [...] Ainsi nous cherchons quelque chose qui est comme un pronom, mais qui occupe une position de phrase. Ce qu'il faut, c'est une *prophrase*. (Grover, Camp et Belnap, 1975, p. 82, italiques de l'auteur)

Or justement, les langues naturelles sont pauvres en prophrases, ce qui nous amène à croire que nous ne disposons pour traduire (1') que de pronoms :

> L'illusion est naturelle précisément parce que l'anglais [N.D.T. ou, dans le cas qui nous concerne, le français], —l'anglais tel qu'il se présente—, ne contient probablement aucune prophrase atomique (en un seul mot) généralement disponible comme le sont « il » et ses cousins. Pour cette raison, une lecture *facile* des énoncés comportant une quantification propositionnelle comme le sont ceux de Ramsey n'est pas disponible. (Grover, Camp et Belnap, 1975, p. 82, italiques de l'auteur)

Selon Grover, Camp et Belnap, la solution de ce problème consiste à réaliser que les expressions « c'est vrai » et « il est vrai » peuvent être analysées et comprises comme des prophrases. Ainsi, l'extrait cité ci-dessus continue comme suit :

> Mais, —et c'est là une des thèses principales de cet article —, l'anglais a bien des prophrases, quoique non atomiques ; nous prétendons que « c'est vai » et « il est vrai » sont des prophrases, et qu'on peut s'appuyer sur ce fait pour répondre à l'objection d' « agrammaticalité » adressée à Ramsey. (Grover, Camp et Belnap, 1975, p. 82)

Pour éclairer un peu plus la nature de cette espèce lexicale « rare » que sont les prophrases, Grover, Camp et Belnap en appellent au phénomène linguistique de l'anaphore [45], dont les pronoms fournissent des exemples paradigmatiques mais non uniques. L'idée générale est la suivante : certains éléments lexicaux de nos langages ont pour fonction principale de permettre des renvois de référence (ce qu'on appelle parfois des

45. En grammaire, une anaphore (à ne pas confondre avec la figure de style) est un mot ou un syntagme qui, dans un énoncé, assure une reprise sémantique d'un précédent segment appelé antécédent. Grover, Camp et Belnap (1975) reconnaissent volontiers ne pas être en capacité de fournir une théorie générale rigoureuse et complète de ce phénomène linguistique. C'est un soin qu'ils laissent aux linguistes. Comme premier pas dans cette direction, ils renvoient aux travaux de Partee (1970). Néanmoins les nombreux exemples qu'ils fournissent donnent, ou sont censés donner, une idée suffisamment claire de ce dont il s'agit. Pour une étude récente du phénomène grammatical de l'anaphore, voyez Apothéloz (1995).

références croisées), que ce soit pour éviter les répétitions fastidieuses, pour dissiper des ambigüités mais également pour permettre des quantifications. Ces items lexicaux prennent alors « anaphoriquement » la place de leurs antécédents. Comme exemple typique de ce phénomène, on trouve bien entendu les pronoms :

Georges voulait acheter une voiture, mais *il* n'a pu s'offrir qu'un vélo.

Dans cet énoncé, le pronom « il » est employé anaphoriquement : il renvoie à son antécédent, en l'occurrence « Georges », dont il prend la place dans la suite de l'énoncé —du moins s'il n'y a pas d'autre personne dans le contexte à laquelle « il » pourrait référer. Le pronom prend donc ici la place d'un syntagme nominal. Pour pouvoir déterminer sa valeur sémantique et celle de l'énoncé dans lequel il apparaît, on doit pouvoir déterminer quel est son antécédent. Un pronom employé anaphoriquement ne peut ainsi pas fonctionner seul.

Ceci étant, les pronoms occupant des postions nominales et dont les antécédents sont en général des groupes nominaux, ne sont pas les seules expressions anaphoriques de la langue ordinaire. Le phénomène de l'anaphore peut également concerner d'autres catégories grammaticales. En français, on peut par exemple trouver des « pro-verbes » tels que « faire » :

Tente de te comporter comme le *fait* le sage, [46]
Agis comme tu penses devoir le *faire*,

mais également des « pro-adjectifs » comme « tel » ou « telle » :

Elle était belle la dernière fois que je l'ai vue, et *telle* je l'ai retrouvée,

et des « pro-adverbes » comme « ainsi » :

Il sautait violemment, et en sautant *ainsi*, se blessa. [47]

46. Cet exemple et les suivants sont empruntés à Rivenc (1998).

47. Nous ne donnons ici que des exemples élémentaires destinés à faire comprendre le mécanisme. Dans la « vraie vie », ou disons dans les langues naturelles telles qu'elles sont effectivement pratiquées, le phénomène de l'anaphore grammaticale peut rapidement devenir d'une complexité redoutable :

— Le tatouage de Monsieur est situé à un endroit que l'honnêteté et la décence m'interdisent de préciser davantage.
— Ah ! bon, mais qu'entendez-vous par là ?
— Oh ! par là, j'entends pas grand-chose !

D'après P. Dac et F. Blanche, *Le Sar Rabindranath Duval*,

« *par là* » joue-t-il ici le rôle d'une prophrase (au moins dans sa première occurrence) ? Et dans sa deuxième occurrence est-il un pronom ou un proadverbe ?

Pour désigner en toute généralité les expressions entrant dans une relation anaphorique telles que les pro-noms, pro-verbes, *etc.* , Grover, Camp et Belnap (1975) avancent le terme de « proforme ».

Parmi les usages anaphoriques de ces proformes, il faut en outre distinguer ce que Grover, Camp et Belnap (1975) eux-mêmes baptisent à la suite de Geach (1967) et Partee (1970), les usages « de paresse » et les usages « quantificationnels ». Dans l'exemple précédent

(a) Georges voulait acheter une voiture, mais *il* n'a pu s'offrir qu'un vélo.

l'antécédent du pronom « il » est explicitement donné. Et, au prix d'une certaine lourdeur, on pourrait très bien se dispenser d'employer ce pronom en répétant le nom auquel il renvoie :

(a') Georges voulait acheter une voiture, mais Georges n'a pu s'offrir qu'un vélo. [48]

On a typiquement à faire ici à un usage de paresse : le recours au pronom est simplement pratique mais pas indispensable. Cependant tous les emplois anaphoriques ne sont pas de cet ordre. Il existe également des emplois quantificationnels, comme l'illustre l'énoncé suivant :

(Q) Tout entier positif est tel que s' *il* est pair, alors lorsqu'on *lui* ajoute 1, on obtient un nombre impair. [49]

Ici, il ne peut être question de remplacer « il » et « lui » par « Tout entier positif » [50], alors même que cette expression quantifiée semble être l'antécédent de ces deux anaphores. Dans les cas d'usages quantificationnels, la relation entre l'anaphore et son antécédent est manifestement significativement différente de celle qui opère dans les cas d'emplois de paresse. Ici, les pronoms n'héritent pas leur référence de l'antécédent. Selon Grover, Camp et Belnap (1975, p. 85), il apparaît plutôt que dans ces cas, l'antécédent déter-

48. Il faudrait sans nul doute nuancer : l'équivalence de (a) et de (a') pourrait n'être pas aussi évidente qu'il y paraît. Par exemple, la présence d'un pronom indique qu'il doit y avoir un antécédent, ce qui est une information que l'expression « Georges » ne contient pas. Toutefois, l'idée devrait être assez claire.

49. Grover, Camp et Belnap (1975, p. 85).

50. « Tout entier positif est tel que si tout entier positif est pair, alors lorsqu'on ajoute 1 à tout entier positif, alors on obtient un nombre impair » ne convient visiblement pas.

mine une famille ou une classe d'expressions substituables [51] pour l'anaphore. Ainsi, une « instance » de (Q) pourrait être

> Si 3 est pair, alors lorsqu'on ajoute 1 à 3, on obtient un nombre impair.

En extrapolant à partir des exemples précédents, ce que sont censées être les « prophrases » devrait être raisonnablement clair : des éléments lexicaux permettant un renvoi anaphorique vers une autre catégorie d'antécédents, à savoir les phrases, ou plus précisément les énoncés déclaratifs, et qui peuvent être employés dans des usages de paresse ou dans des usages quantificationnels. Voici la façon dont Grover, Camp et Belnap (1975) eux-mêmes proposent de les caractériser :

> Nous rassemblons ces considérations pour aboutir à l'ébauche d'un critère concernant ce qu'on peut bien vouloir dire par « prophrase » :
>
> (i) Elle peut occuper la position d'un énoncé déclaratif.
>
> (ii) Elle peut être employée anaphoriquement, que ce soit dans les usages de paresse ou dans les usages quantificationnels.
>
> (iii) Par conséquent, dans chacun de ces emplois, elle possède un antécédent à partir duquel on peut dériver un *subsituend* [52] anaphorique (dans les cas de paresse) ou une famille de *substituends* anaphoriques (dans les cas quantificationnels) —dans chaque cas, les *substituends* sont des phrases, correspondant à la position de l'anaphore.
>
> (iv) Elle est « générique » au sens où, dans l'un ou l'autre de ces emplois, n'importe quel énoncé déclaratif pourrait être convoqué comme *substituend* anaphorique.
>
> (Grover, Camp et Belnap, 1975, p. 87)

Selon Grover, Camp et Belnap, l'anglais (tout comme le français d'ailleurs) ne dispose pas de prophrases « atomiques », c'est-à-dire de termes ou d'expressions formées

51. Ou peut-être faudrait-il dire une classe d'*objets*. Mais, précisément, Grover, Camp et Belnap (1975) veulent éviter de traiter ces quantifications anaphoriques de manière objectuelle. D'ailleurs rien n'exclut à leurs yeux la possibilité d'usages quantificationnels d'anaphores autres que pro-nominales, comme par exemple des anaphores pro-verbales, ou pro-adjectivales, ou pro-adverbiales, *etc.* , même si celles-ci semblent rares, voire introuvables, en anglais. Dans de tels cas, la classe de substitution ne contiendrait pas des (noms d') objets, mais des verbes, des adjectifs, des adverbes, *etc.*

52. N.D.T. nous conservons ici le latinisme, quoi qu'il soit moins courant en français qu'en anglais. Ce terme désigne en logique et en linguistique : « une chose qui peut être mise à la place d'une autre ; un mot, une expression, *etc.* qui peut se substituer à un autre » (d'après la définition du Oxford dictionary).

d'un seul mot qui satisferaient les conditions (i) à (iv) ci-dessus.[53] Ils suggèrent alors de considérer les locutions « c'est vrai » et « il (elle) est vrai(e) » comme des prophrases. De ce point de vue, la syntaxe de surface de ces tournures, c'est-à-dire leur structure grammaticale apparente de la forme « sujet verbe prédicat », est trompeuse et rend difficile à percevoir leur rôle de prophrases. Il faudrait au contraire les traiter comme des atomes sémantiques, ou bien encore, si l'on préfère, comme des éléments lexicaux indécomposables. On pourrait peut-être rendre ceci plus apparent en les notant « C'est-vrai » et « Il-est-vrai ».

Pour insister sur ce caractère indécomposable de ces prophrases « aléthiques[54] » Grover, Camp et Belnap (1975) proposent en parallèle de munir l'anglais d'un nouveau terme « *thatt* » désignant un opérateur de prophrase. On peut imiter cette construction en enrichissant le français d'un nouvel élément lexical que nous appellerons « *celaa*[55] ». L'idée défendue par Grover, Camp et Belnap (1975) est que les prophrases aléthiques ou la prophrase « celaa » ont un comportement sémantique rigoureusement identique : elles servent de support à des anaphores grammaticales portant sur la catégorie lexicale des énoncés.

Dans un langage muni de ces prophrases, les constructions ramseyiennes impliquant des variables propositionnelles deviennent irréprochables sur un plan grammatical :

(1) Tout ce que dit Jean est vrai

53. Ils reconnaissent toutefois que les langues naturelles possèdent bien des mots qui peuvent parfois jouer le rôle de prophrases dans certaines constructions particulières, notamment : « yes », « oui », ou certains usages de « so » en anglais. Brentano (1930) déjà appelait le mot « oui » (« Ja ») un « Fürwort », c'est-à-dire littéralement un « pro-mot ». À titre d'illustration considérez la phrase :

Je ne crois pas que Georges soit malade, mais si *oui*, il doit rester chez lui

même si cette construction peut sonner bizarrement en français, « oui » semble bien jouer ici un rôle de prophrase (l'anglais dirait « if *so* », dans un français plus soutenu on emploierait peut-être plutôt la locution « tel est le cas » qui est une prophrase non-atomique).

Quoi qu'il en soit, comme Grover, Camp et Belnap (1975) le soulignent, ces termes ne peuvent jouer le rôle de prophrases de circonstance que dans des cas d'usages de paresse. Les usages quantificationnels restent hors de portée. On ne peut pas dire :

Pour toute proposition si Jean dit que *oui*, alors *oui*

54. Aléthiques au sens où le prédicat de vérité *semble* y figurer quoiqu'il n'y ait en réalité aucune prédication.

55. Rivenc (1998, p. 32) propose pour sa part d'introduire en français l'opérateur « Cela » (avec un *C* majuscule).

que Ramsey rendait par

$$(1') \ \forall p \ (\text{Jean dit que } p \rightarrow p)$$

se paraphrase au moyen des prophrases par

(1*) Pour toute proposition, si Jean dit qu'*elle-est-vraie*, alors *elle-est-vraie*.

(1**) Pour toute proposition, si Jean dit que *celaa*, alors *celaa*.

Un autre exemple d'usage quantificationnel de ces prophrases peut s'obtenir à partir du principe de bivalence suivant :

(2) Toute proposition est soit vraie soit fausse.

qui donne « à la Ramsey »

$$(2') \ \forall p(p \vee \neg p)$$

et est para(pro)phrasable par

(2*) Pour toute proposition, soit *elle-est-vraie*, soit *elle-n'-est-pas-vraie*.[56]

(2**) Pour toute proposition, soit *celaa*, soit non *celaa*.

Illustrons aussi ce que donne un usage de paresse :

(3*) Il y a de la vie sur Mars ! Si *c'est-vrai*, alors nous devrions bientôt observer des signes de vie.

(3**) Il y a de la vie sur Mars ! Si *celaa*, alors nous devrions bientôt observer des signes de vie.

Il est important de bien saisir ici le comportement sémantique des prophrases « il-est-vrai » ou « c'est-vrai ». Dans les constructions ci-dessus la relation d'anaphore et le renvoi de référence qui s'en suit ne sont pas portés par les pronoms *il* ou *c'* mais bien par la prophrase complète et, répétons-le, indécomposable. C'est ici que la mise en parallèle avec le terme « celaa » est particulièrement éclairante. Plus précisément, dans (3*)/(3**) par exemple, l'antécédent « Il y a de la vie sur Mars » se substitue à toute la prophrase « c'est-vrai »/ « celaa » et non pas au pronom « c' ». Ainsi, après substitution dans (3*), on obtiendrait

56. Attention, là encore la grammaire de surface est trompeuse : la négation doit porter sur la prophrase, *i.e.* « elle-est-vraie », et non pas sur le prédicat « vrai » (mais quel prédicat ? la prophrase « elle-est-vraie » est un atome sémantique indécomposable...). Dans un langage semi formel, on pourrait le noter ¬(elle-est-vraie).

S'il y a de la vie sur Mars, alors ... [57],
et non pas : Si « il y a de la vie sur Mars » est vrai, alors .... [58]

De même, et plus crucialement, une instance de l'énoncé quantifié

(1*) Pour toute proposition, si Jean dit qu'*elle-est-vraie*, alors *elle-est-vraie*.

sera

(1_i) Si Jean dit que la lune est verte, alors alors la lune est verte

où un énoncé a remplacé la prophrase, et non pas

Si Jean dit que « la lune est verte » est vrai , alors alors « la lune est verte » est vrai

où c'est le pronom qui a été remplacé par un nom d'énoncé obtenu par mise entre guillemets.

Il est sans doute utile de souligner ici la conception de la quantification qui semble accompagner l'emploi des prophrases. Si (1_i) est une instance de (1*), on peut remarquer qu'on l'obtient non pas en remplaçant l'élément porteur de l'anaphore (*i.e.* l'élément variable portant un renvoi de référence, c'est-à-dire ici la prophrase [59]) par un objet, mais en lui *substituant* un syntagme linguistique, à savoir une phrase assertive. Dès lors, la quantification semble porter non pas sur un domaine d'objets, mais bien plutôt sur une classe d'éléments syntaxiques substituables, en l'occurrence des énoncés. Ceci suggère dont très fortement une interprétation substitutionnelle de la quantification. Ce type d'interprétation de la quantifiction est d'ailleurs explicitement mentionné par Grover, Camp et Belnap (1975), et plusieurs passages laissent à penser que c'est bien cette conception que les auteurs ont en tête. Par exemple, ils affirment que la prophrase

> « celaa [N.D.T. *thatt*] » n'est *jamais* une expression référentielle, qu'elle soit utilisée quantificationnellement ou comme une prophrase de paresse. [...] il n'est pas du tout facile de donner un sens intuitif à une question comme « Qu'est-ce que *parcourt* « celaa » dans « Toute proposition est telle que soit celaa, soit non celaa » ? » [...] les questions concernant ce que parcourt « celaa » sont déplacées, comme le sont les théories de l'engagement ontologique qui assumeraient que « celaa » doit « parcourir » quelque chose. (Grover, Camp et Belnap, 1975, p. 89-90)

57. où « il y a de la vie sur Mars » se substitue à « c'est-vrai ».
58. où un noms d'objet, à savoir « « il y a de la vie sur Mars » » se substitue à « c' ».
59. et que dans une langage formel on rendrait par une variable propositionnelle.

De même, une interprétation substitutionnelle est explicitement mise en avant quelques pages plus loin

> [...] nous sommes plutôt favorables à la perspective d'une sémantique qui conçoit [« Toute chose est telle que si Charley croit qu"elle est vraie, alors elle est vraie »] comme, en fait, une quantification substitutionelle, où la vérité de [cet énoncé] est équivalente à la vérité de toutes ses instances de substitution (« Si Charley croit que la neige est blanche alors la neige est blanche », *etc.*). (Grover, Camp et Belnap, 1975, p. 113)

Plus généralement, la théorie prophrastique a, la plupart du temps, été comprise comme *devant* s'accompagner d'une interprétation substitutionnelle de la quantification [60]. Faut-il y voir une limite et un problème pour la théorie prophrastique de la vérité ? Peut-être. Il est vrai que le statut de la quantification susbtitutionnelle est pour le moins controversé et n'admet pas de réponse philosophique qui aille de soi. Toutefois, comme ce serait un sujet à part entière de discuter ce point en profondeur, nous laisserons cette question de côté [61]. Attardons-nous plutôt sur les leçons philosophiques que Grover, Camp et Belnap (1975) entendent tirer de la théorie prophrastique de la vérité.

Tout d'abord, remarquons que Grover, Camp et Belnap (1975) ne se contentent pas d'affirmer que « il est vrai » et « c'est vrai » peuvent (et doivent) être analysés comme des prophrases. Bien au contraire, ils souscrivent à une thèse beaucoup plus forte selon laquelle tous les « parlers » en termes de vérité, tout ce que l'on peut dire en employant cette notion, y compris lorsqu'elle semble apparaître sous une forme prédicative, peut être réduit et reformulé en un discours qui n'emploiera *que* des prophrases.

> Dans l'esprit de Ramsey, notre thèse est que *tout* le parler en termes de vérité peut être vu comme n'impliquant que des emplois prophrastiques de « c'est vrai ». (Grover, Camp et Belnap, 1975, p. 92, italiques des auteurs)

60. Sur ce point, voyez Rivenc (1998). Horwich (1998b, p. 26 et 125) en particulier semble penser que la théorie prophrastique de la vérité va nécessairement de pair avec une lecture substitutionnelle de la quantification. Indiquons néanmoins que Kirkham (1992, p. 329-335) affirme quant à lui que la quantification substitutionnelle n'est qu'une des « stratégies » possibles pour résoudre le problème des variables prophrastiques.

61. Nous renvoyons le lecteur à l'abondante littérature portant sur cette question : Dunn et Belnap (1968) peut constituer un point d'entrée, le classique Kripke (1976) propose un examen approfondi (très) critique des problèmes soulevés par cette conception de la quantification. Signalons égalemnet Shapiro (2001, p. 180-187) qui s'attarde sur les difficultés techniques posées par cette interprétation des quantificateurs : en particulier, la sémantique substitutionnelle n'est pas compacte et, selon les conventions adoptées, elle pourra parfois être également ineffective. Pour une discussion en lien avec la théorie prophrastique ainsi que des références supplémentaires, voyez à nouveau Rivenc (1998).

Pour défendre leur proposition, ils introduisent l'anglais* [62] une variante de l'anglais expurgée du prédicat de vérité et qui ne contient que les prophrases « it is true » et « that is true » [63]. Comme nous l'avons déjà expliqué, ces prophrases sont à prendre comme des atomes sémantiques. Autrement dit, l'anglais* ne contient pas de *prédicat* de vérité. Les auteurs s'attachent ensuite à montrer comment on peut reformuler n'importe quel énoncé de l'anglais en un énoncé de l'anglais*. Dès lors, disent-ils,

> Nous pouvons à présent énoncer rapidement l'une des principales thèses de notre théorie prophrastique de la vérité : l'anglais peut se traduire *sans perte significative* dans son sous-fragment l'anglais*. Et une thèse supplémentaire est qu'une telle traduction est *claire et explicative*. (Grover, Camp et Belnap, 1975, p. 93, nous soulignons)

Une fois établie la possibilité d'une telle traduction, Grover, Camp et Belnap en concluent :

> Ainsi, le parler en termes de vérité de l'anglais est sémantiquement et pragmatiquement semblable au parler en terme de vérité de l'anglais*. Les deux langues possèdent des conventions grammaticales différentes, mais en anglais, tout comme en anglais*, le prédicat de vérité ne joue pas le rôle d'un attributeur de propriété. La vérité, pour le dire en une formule, *n'est pas un véritable prédicat*. (Grover, Camp et Belnap, 1975, p. 97, nous soulignons)

> Notre suggestion est que les prophrases quantificationnelles sont de ce point de vue semblables aux pronoms quantificationnels : elles sont absolument non redondantes en anglais*, nous permettant de dire des choses que l'on ne pourrait pas dire sans elles, mais cette *non-redondance est de nature logique*, comme celle de « ou ». Et, si l'on examine l'autre côté de la médaille, elles sont (contrairement à ceux qui considèrent la vérité comme un attribut) thématiquement [*topically*], catégoriquement [*categorically*] et —tout particulièrement— attributivement [*ascriptively*] redondantes. [64] (Grover,

62. La même démarche serait évidemment transposable au français mais nous garderons ici l'exemple de la langue d'origine de Grover, Camp et Belnap (1975).

63. Pour être tout à fait précis, Grover, Camp et Belnap (1975) recommandent d'ajouter aux prophrases « it is true [il est vrai] » et « that is true [c'est vrai] », une série d'autres opérateurs permettant de traiter les variations de temps ou de modalité, tels que « il était vrai que », « il sera vrai que », « il est nécessairement vrai que », ... Nous n'entrerons pas dans ces complications techniques. Voyez Grover, Camp et Belnap (1975, p. 93) pour plus de précisions et une liste plus développée de ces autres opérateurs.

64.

Camp et Belnap, 1975, p. 123, nous soulignons)

Pour un auteur sans doute non-déflationniste lui-même, Ramsey aura eu une descendance déflationniste pour le moins fertile ! Le trait saillant des analyses ci-dessus est avant tout le refus de considérer la vérité comme une authentique propriété. Ces divers auteurs remettent en cause la structure prédicative apparente du mot « vrai », et proposent des reformulations et des schématisations du langage naturel où le prédicat de vérité est, soit explicitement éliminé, soit dénué de tout rôle prédicatif. Ces théories font souvent appel à la quantification substitutionnelle.

À travers les réflexions d'un Frege ou d'un Ramsey et de ses héritiers, nous avons donc pu voir apparaître divers thèmes touchant la vérité : l'absence de contenu propre du prédicat de vérité ; la transparence des attributions de vérité et la trivialité de l'équivalence entre une proposition/un énoncé/une croyance que p et la proposition/l'énoncé/la croyance que p est vrai ; les premiers doutes concernant le statut d'authentique propriété de la notion de vérité ; les difficultés soulevées par la fonction du prédicat de vérité dans l'expression de certaines généralisations et ses liens avec la quantification. Dans leurs versions les plus radicales, ces réflexions ont pu aboutir à des théories éliminatives de la vérité selon lesquelles le prédicat de vérité est simplement redondant. Une conséquence importante de ce type de conception est que la vérité ne peut jouer de rôle important et fondamental dans nos explications du monde puisque celles-ci pourront en principe se formuler sans recourir à cette notion. Nous allons voir que ces thèmes ont été repris, transformés et redéveloppés par les auteurs déflationnistes contemporains. Mais avant d'en venir à ces auteurs proprement dits, il nous faut encore faire un détour par un autre précurseur dont l'influence sur les déflationnistes a été prépondérante.

Ceci suggère, nous supposons, que les pronoms quantificationnels nous offrent des façons de dire certaines choses en anglais qu'autrement nous ne pourrions pas dire. Mais notez que les choses nouvelles que nous pouvons dire à l'aide des pronoms quantificationnels et que nous ne pouvons pas dire sans eux, sont nouvelles en un sens particulier : elles ne sont ni *thématiquement* [*topically*] nouvelles, elles ne nous permettent pas d'aborder de nouveaux thèmes de discussion ; ni *attributivement* [*ascriptively*] nouvelles, elles ne nous fournissent aucune nouvelles propriétés ou nouvelles relations indécomposables ; ni *catégoriquement* [*categorically*] nouvelles, elles ne nous procurent pas un cadre conceptuel flambant neuf à l'intérieur duquel travailler. La nouveauté des choses que nous pouvons dire est plutôt semblable à celle des choses que nous pouvons dire avec « ou » et que nous ne pouvons pas dire sans « ou » ; disons que cette nouveauté est *logique*. (Grover, Camp et Belnap, 1975, p. 123)

1.1.4 La théorie de Tarski

Dans cette section, nous voudrions exposer les principaux résultats de Tarski. Ceci paraît en effet indispensable dans la mesure où les travaux de Tarski sur le concept de vérité ont exercé une influence décisive sur pratiquement tous les auteurs et théoriciens de la vérité, philosophes [65] ou logiciens, qui sont venus après lui. Nous allons voir qu'il a notamment été une source d'inspiration primordiale pour les penseurs déflationnistes.

Ses résultats et réflexions ont pour la plupart été formulés par Tarski dans ce qui constitue sans doute l'article le plus fameux concernant les théories formelles de la vérité : *Le concept de vérité dans les langages formalisés*. Rappelons que le destin éditorial de ce célèbre texte fut assez tortueux : Tarski avait d'abord rédigé et publié une première version en polonais (Tarski, 1933) [66]. Quelques années plus tard, paraissait une traduction en allemand (Tarski, 1935) [67]. Elle était accompagnée d'un important *Post-Scriptum* dans lequel l'auteur revenait sur certaines de ces conceptions [68]. Une traduction anglaise, approuvée par Tarski lui-même et à l'occasion de laquelle il ajouta également quelques notes, est ensuite parue dans les années cinquante (*in* Tarski (1956a)). Enfin, une traduction française établie à partir des diverses versions (polonaise, allemande et anglaise) et réalisée sous la direction de G. G. Granger fut publiée dans les années 1970 (*in* Tarski (1972)). Les articles Tarski (1944, 1969) dans lesquels Tarski reprend les principales conlusions philosophiques de son travail tout en laissant de côté les aspects logico-mathématiques les plus techniques, étaient quant à eux destinés à un plus large public.

Pour saisir pleinement le sens et les objectifs poursuivis par Tarski dans son travail sur la vérité, il faut le replacer quelque peu dans le contexte historique qui l'a vu naître. Au

65. Du moins pour ceux qui sont rattachés à la tradition analytique.

66. Avant même cette parution, les travaux de Tarski avaient été présentés (par J. Łukasiewicz) à la société des Lettres et des Sciences de Varsovie. Tarski (Tarski (1935, p. 161, note 2 ; la pagination renvoie à la traduction française dirigée par G.G. Granger) indique qu'il avait obtenu une large part de ses résultats dès 1929 et que ceux-ci avaient fait l'objet de deux communications préalables. Un résumé de ces résultats était paru en polonais dès 1931 (Tarski, 1930-31), tandis qu'un autre résumé (Tarski, 1932), en allemand celui-là et qui portait sur l'ensemble de l'étude, était également paru en 1932, c'est-à-dire là encore avant la (première) publication (en polonais) du texte intégral de la monographie. Pour plus de précisions, nous renvoyons à Tarski (1935, p. 161, note de bas de page), ou bien encore à Givant (1986).

67. Cette traduction, parue en tiré à part en 1935, fut conjointement publiée dans le volume I de *Studia Philosophica* en 1936.

68. Sur l'importance de ce *Post-Scriptum* et sur le fait qu'il ne constitue pas un simple prolongement ou une simple modification à la marge, mais bel et bien un renversement important des conceptions philosophiques de l'auteur, voyez Rouilhan (1998).

tournant des années trente, nombreux étaient les philosophes qui portaient un regard plein de méfiance envers le concept de vérité, et plus généralement envers l'ensemble des notions relevant de la sémantique[69]. Outre l'existence de conceptions apparemment irréconciliables regardant la notion de vérité —conception de la vérité correspondance, conception de la vérité cohérence, conception pragmatiste de la vérité— qu'on pouvait interpréter comme le symptôme d'une irrémédiable obscurité de ce concept, l'une des principales source d'inquiétude provenait de la (re)découverte des nombreux paradoxes auxquels les notions sémantiques semblaient conduire. Tarski lui-même caractérisait la situation comme ceci :

> Depuis l'Antiquité jusqu'à nos jours les concepts sémantiques on joué un rôle important dans les discussions des philosophes, des logiciens et des philologues. Néanmoins ces concepts ont été longtemps traités avec une certaine méfiance. Du point de vue historique, cette méfiance doit être reconnue comme entièrement justifiée. Car bien que le sens des concepts sémantiques tels qu'ils sont employés dans le langage quotidien semble être plutôt clair et intelligible, toutes les tentatives entreprises en vue de la caractérisation de ce sens, d'une manière à la fois générale et exacte échouèrent. Qui plus est, diverses argumentations comportant ces concepts et qui par ailleurs semblaient tout à fait correctes et fondées sur des prémisses apparemment évidentes conduisirent fréquemment à des paradoxes et à des antinomies. Il suffit de mentionner ici l'*antinomie du menteur*, l'*antinomie de la définissabilité* de Richard (où il s'agit de définition contenant un nombre fini de mots) et l'*antinomie des termes hétérologiques* de Grelling-Nelson.
>
> (Tarski, 1944, p. 274, italiques de l'auteur, la pagination renvoie à la traduction française) [70]

69. C'était le cas en particulier des membres du Cercle de Vienne avec lesquels Tarski était en contact dès les années trente. Sur les débats concernant la vérité animant le Cercle à cette époque et sur les réactions (contrastées) et les résistances que les travaux de Tarski suscitèrent en son sein, voir Mancosu (2015b). Concernant les discussions à propos des notions sémantiques en cours au coeur de l'école polonaise, dite de Lvov-Varsovie, dont Tarski était plus directement issu, voir Simons et Woleński (1989) et Vuissoz (1998), ainsi que Patterson (2012, 2008).

70. Pour les extraits de Tarski (1935) et Tarski (1944) cités ici, nous reprenons la traduction française parue sous la direction de G. G. Granger Tarski (1972, 1974), à quelques modifications près. En particulier, nous avons traduit l'expression originale « sentence/Aussagen » par « énoncé » plutôt que par « proposition ». Nous suivons sur ce point l'usage contemporain de la littérature logique et philosophique francophone, qui consiste à réserver le terme « proposition » pour désigner la signification ou l'intension des énoncés (quel que soit ce qu'on entend par là) plutôt que les énoncés ou les phrases eux-mêmes.

Face à ces difficultés, l'objectif principal de Tarski était donc de fournir une analyse de la notion de la vérité suffisamment claire et logiquement impeccable pour montrer qu'un usage cohérent de ce concept, compatible avec les canons d'une stricte méthodologie scientifique, était possible [71]. Idéalement, lorsque cela est possible, une telle clarification devait prendre la forme d'une définition explicite dont les termes employés —les termes apparaissant dans le *definiens*— seraient non problématiques.

Ainsi, Tarski introduit son mémoire de 1935 par ces mots :

> Le présent travail est consacré presque exclusivement à un seul problème, au *problème de la définition de la vérité*. Il s'agit en effet — compte tenu de tel et tel langage — de *construire une définition de l'expression « énoncé vrai », définition qui soit matériellement adéquate et formellement correcte.*
>
> (Tarski, 1935, p. 159, italiques de l'auteur)

Il poursuit ensuite :

> Le problème de la définition d'un concept quelconque n'est pas convenablement posé aussi longtemps que n'est pas dressée la liste des termes au moyen desquels on se propose de construire la définition recherchée. En outre, pour que cette définition atteigne son but, le sens des termes figurant dans la liste

71. C'est du moins ce qu'en a retenu l'interprétation traditionnelle. Popper (2005, p. 273, note de bas de page) est une illustration classique de ce type de réception des travaux de Tarski. Voyez également l'« Autobiographie » de Carnap dans Schilp (1963, p. 61) pour un autre témoignage concernant l'accueil, la compréhension et l'influence des travaux tarskiens.

Toutefois, que le lecteur soit averti : la question de savoir quels étaient les objectifs *implicites* ou *explicites* poursuivis par Tarski, tout comme celle de savoir ceux qu'il a pleinement atteints ou non, est aujourd'hui encore l'objet d'intenses discussions parmi les spécialistes. Par exemple, Kirkham (1992, chapitre 5, p. 141-144) attribue à Tarski une grande variété d'objectifs ou de « programmes » dans son travail sur la vérité :

- fournir des fondations sûres à ce que Tarski lui-même appelait la « sémantique scientifique » ;
- sous l'influence du physicalisme, ne laisser aucun terme sémantique indéfini (ou primitif) et éviter toute circularité dans les définitions des termes sémantiques, mais les réduire tous à des termes acceptables aux yeux d'un « physicaliste » ;
- montrer comment la structure grammaticale des énoncés influence leur valeur de vérité, ou pour reprendre la formulation de Kirkham : créer une « théorie des modèles » pour la logique des prédicats quantifiés, à l'image de celle qui existait déjà à l'époque pour la logique propositionnelle ;
- se garantir contre les paradoxes tels le menteur ;
- mettre au point une théorie de la vérité qui satisfasse le critère d'adéquation matérielle.

Plus récemment encore, l'étude historique de Patterson (2012) a proposé une nouvelle lecture de l'œuvre de Tarski sur la vérité. Selon Patterson, l'entreprise de Tarski ne peut se comprendre qu'à la lumière du contexte philosophique de l'Ecole de Lvov-Varsovie dont il était issu. La thèse centrale de Patterson est que Tarski ne se serait dépris de l'influence de ses maîtres polonais (en particulier de Lesniewski et de son « formalisme intuitionniste ») que très progressivement et très tardivement au cours des années trente. Cette influence était donc encore prégnante lors du développement et de la mise au point de ses travaux sur la vérité.

> en question doit être indubitable. [...] je n'ai l'intention d'utiliser pour cette construction aucun concept sémantique qui ne soit pas antérieurement réduit à quelque autre concept. (Tarski, 1935, p. 159-160)

Pour mener à bien son projet, Tarski doit donc, premièrement, préciser aussi rigoureusement que possible les conditions d'adéquation qu'une telle définition devra remplir, c'est-à-dire établir à quelles conditions la définition en question pourra être considérée comme une définition correcte ayant effectivement saisi le concept visé ; deuxièmement, spécifier à quelles conditions une telle définition est possible ; et enfin montrer le cas échéant comment une telle définition s'obtient. Nous suivons les grandes lignes de ces trois temps dans la suite de notre exposé.

1.1.4.1 Critères d'adéquation et de correction

Concernant la première étape, c'est-à-dire la recherche d'un critère d'adéquation, il faut se souvenir que la définition recherchée par Tarski n'est pas une définition purement stipulative, qui fixerait plus ou moins arbitrairement les conditions d'emploi d'un terme nouvellement introduit. Bien au contraire, Tarski entend fournir une définition de la notion ordinaire de vérité, telle qu'elle est apparemment employée dans les sciences ou le discours commun. Autrement dit, Tarski considère la vérité comme une notion préthéorique et intuitive, qu'il s'agit de préciser et de définir adéquatement :

> La définition souhaitée ne vise pas la détermination du sens d'un mot familier employé pour signifier une notion nouvelle ; elle voudrait, au contraire, saisir le sens effectif d'une vieille notion. Nous devons donc caractériser cette notion de manière suffisante pour permettre à chacun de constater si la définition remplit effectivement cette tâche. (Tarski, 1944, p. 269)

Plus précisément, pour Tarski, la définition recherchée devra capturer la signification, ou du moins l'extension, de la notion de vérité telle qu'elle est conçue, selon lui, dans la conception classique de la vérité comme correspondance :

> Nous voudrions que notre définition rendît justice aux intuitions qui sont celles de la *conception classique aristotélicienne de la vérité*, intuitions qui trouvent leur expression dans cette phrase bien connue de la *Métaphysique* d'Aristote :
>
> *Dire de ce qui est qu'il n'est pas ou de ce qui n'est pas qu'il est est*

> *faux tandis que dire de ce qui est qu'il est, et de ce qui n'est pas qu'il n'est pas est vrai.*[72]

Si nous désirons nous conformer à la terminologie philosophique moderne, nous pourrions peut-être exprimer cette conception au moyen de la formule bien connue :

> *La vérité d'un énoncé consiste en son accord (ou sa correspondance) avec la réalité.*

(On a suggéré pour la théorie basée sur cette dernière formule la nom de « théorie de la correspondance »). (Tarski, 1944, p. 270, italiques de l'auteur)

Tarski baptisera cette conception « classique » de la vérité, précisée et revisitée par ses soins, la *conception sémantique* de la vérité. Afin de caractériser plus exactement cette notion sémantique de vérité mise en avant par Tarski, un certain nombre de clarifications préalables sont nécessaires.

Tout d'abord, il faut spécifier la nature des porteurs de vérité et le type de langages concernés. Sur la question des porteurs de vérité, Tarski est le digne héritier des membres de l'école Lvov-Varsovie[73] : les porteurs de vérité seront les énoncés d'un langage identifiés selon leurs types ; c'est-à-dire que le prédicat « (est) vrai » s'appliquera non pas à des énoncés considérés comme des objets physiques concrets individuels, comme telle ou telle suite réelle de graphèmes ou de phonèmes (*i.e.* tel ou tel *token*), mais plutôt à des objets abstraits qui constituent la forme générale de ces inscriptions (*i.e.* un *type*) :

> Il est quelque peu préférable pour notre propos d'entendre par « expression », « énoncés », *etc.*, non les inscriptions individuelles, mais des classes d'inscriptions homéomorphes (donc non des choses physiques mais des classes de ces choses)[74]. (Tarski, 1944, p. 270, note de bas de page)

D'autre part, Tarski insiste également sur le fait qu'il n'y a de sens à appliquer la vérité qu'à des énoncés *interprétés*, c'est-à-dire à des énoncés considérés comme faisant

72. Aristote (*Metaphysique*, Γ, 7, 27, traduction Tricot).

73. Sur le développement de la sémantique au sein de l'Ecole polonaise jusqu'à Tarski et plus particulièrement sur la question des porteurs de vérité, voyez Rojszczak (2005).

74. Voyez aussi Tarski (1935, note 3 p. 163) pour un développement similaire. Notons qu'en renonçant aux énoncés considérés uniquement comme des inscriptions physiques, Tarski s'éloigne du nominalisme strict de son professeur Lesniewski. Sur le nominalisme draconien de ce dernier, voyez Simons (1996) et surtout Simons (2002). Sur le nominalisme propre à Tarski, auquel celui-ci restera fidèle toute sa carrière, et sur les rapports de ce nominalisme avec celui de Lesniewski, voyez Simons (2008) et Patterson (2012). Sur les affinités du nominalisme tarskien avec le nominalisme qui fut un temps celui de Quine voyez Mancosu (2015a) et Frost-Arnold (2008).

partie d'un langage fixé et dont la « signification intuitive » est déjà donnée.

> Il paraît donc préférable [...] d'appliquer le terme « vrai » aux énoncés et nous procéderons ainsi.
>
> En conséquence, nous devons rapporter toujours la notion de vérité, tout comme celle d'énoncé, à un langage déterminé ; car il est évident que la même expression, énoncé vrai dans un langage, peut être un énoncé faux dans un autre. (Tarski, 1944, p. 270)

De même,

> [...] nous ne nous intéressons point aux langages et aux sciences « formels » dans un sens particulier de ce terme, notamment aux sciences dont les signes et expressions ne se voient attribuer aucun sens intuitif. Par rapport à ces sciences le problème posé ici [N.D.T. celui de donner une définition de la vérité] perd toute sa raison d'être et cesse tout simplement d'être intelligible. Aux signes concernés par les présentes considérations nous attribuons toujours une signification tout à fait concrète et intelligible [...] (Tarski, 1935, p. 173)

Ainsi, les langages pour lesquels Tarski entend définir la vérité ne sont pas considérés comme de simples suites de combinaisons de symboles dénuées de signification, mais bien comme des langages interprétés et déjà dotés d'une « signification intuitive ». Pour Tarski, il n'y a donc pas de problème à considérer un langage qui soit simultanément formel et interprété. Sur la nature de cette « signification intuitive », sur la manière dont nous y avons accès, Tarski reste peu disert. Toutefois, il est clair qu'une compréhension préalable du signe est à ses yeux une condition *sine qua non* à la sémantique[75]. Dans la mesure où un même énoncé considéré d'un strict point de vue syntaxique peut avoir

75. Sur ce point, voyez Patterson (2009). Ici, la position de Tarski est sans doute à rapprocher de celle de son maître Lesniewski, parfois qualifiée de « formalisme intuitionniste ». Lesniewski caractérisait de la manière suivante sa propre position consistant à prendre comme constituants des théories des énoncés formalisés *et* interprétés :

> N'ayant aucune prédilection pour les « jeux mathématiques variés » qui consistent à écrire selon une règle conventionnelle ou une autre des formules plus ou moins imagées qui n'ont pas besoin d'avoir de signification et même — comme préfèrent peut-être quelques « joueurs mathématiciens »— doivent être dénuées de signification, je n'aurais pas pris la peine de systématiser et de vérifier souvent si scrupuleusement les directives de mon système, si je n'avais imputé à ses thèses une certaine signification spécifique et complètement déterminée, en vertu de laquelle ses axiomes, définitions et directives finales [...] ont pour moi une validité intuitive irrésistible. (Lesniewski (1929), cité dans Woleński (2009) p. 49)

telle signification dans tel langage interprété et telle autre signification dans tel autre langage interprété, il faudra toujours préciser pour quel langage interprété la définition de la vérité est donnée. Autrement dit, ce que Tarski cherche à définir c'est un prédicat du type « vrai-dans-$\mathcal{L}$ » pour un langage (interprété) $\mathcal{L}$ fixé.

Signalons enfin un dernier point, mais non des moindres, à propos des langages tels qu'envisagés par Tarski : lorsqu'il examine un langage pour lequel il souhaite définir un prédicat de vérité, non seulement Tarski le prend comme un langage interprété, mais en plus il l'accompagne aussitôt, pour ainsi dire dans le même mouvement, d'un ensemble d'axiomes exprimés dans ce langage et de règles gouvernant les inférences donnant des conditions d'assertabilité pour les énoncés de ce langage. Autrement dit, conformément à un usage répandu à son époque [76], ce que Tarski appelle langage est constitué non seulement de ce que nous nommerions aujourd'hui un langage *stricto sensu*, c'est-à-dire un ensemble d'expressions primitives doté de règles de formation syntaxique et éventuellement muni d'une interprétation, ***mais également*** de ce que nous désignerions comme une théorie exprimée dans ce langage, c'est-à-dire *grosso modo* un ensemble d'axiomes adossés à un système déductif permettant d'en dériver des théorèmes. Pour Tarski, les langages pour les énoncés desquels il s'agit de définir un prédicat de vérité sont toujours des langages interprétés et des langages (munis) d'une théorie.

Ces précisions apportées, voyons comment Tarski poursuit sa caractérisation de la notion sémantique de vérité et quelles sont les conditions d'adéquation qu'une définition satisfaisante de cette notion devra, selon lui, remplir.

Nous avons vu que Tarski se réclamait d'une conception classique et correspondantiste de la vérité qu'il faisait remonter à Aristote. Cependant, les formulations qui en ont été données jusqu'à présent ne lui semblent guère satisfaisantes et ne peuvent en tout cas pas être prises comme des définitions. Il faut donc, nous dit-il, chercher une formulation plus précise de nos intuitions [77]. Pour ce faire, Tarski se tourne vers l'examen d'un exemple concret [78] : prenons l'énoncé « la neige est blanche ». À quelles conditions peut-on dire qu'il est vrai ? Trivialement, il sera vrai si est seulement si la neige est blanche. C'est de cette simple et banale observation qu'il faut partir, et Tarski écrit :

> [...] si notre définition de la vérité doit être conforme à notre conception de celle-ci, elle doit impliquer l'équivalence suivante :

76. mais qui n'a plus cours à présent.
77. *Cf.* Tarski (1944, p. 271)
78. devenu depuis particulièrement célèbre.

> *L'énoncé « la neige est blanche » est vrai si et seulement si la neige est blanche.*

(Tarski, 1944, p. 271)

Pour Tarski, l'intuition correspondantiste est clairement présente dans ce biconditionnel : celui-ci affirme qu'une certaine entité linguistique, à savoir un énoncé désigné par un nom obtenu par une mise entre guillemets, possède une certaine propriété, à savoir la vérité, si et seulement si un certain état de chose est réalisé. Tarksi va jusqu'à affirmer que cette équivalence peut être considérée comme « une définition partielle » de la vérité, définition partielle qui explique en quoi consiste la vérité de *cet* énoncé particulier qu'est « la neige est blanche ». Bien entendu, on peut généraliser l'exemple ci-dessus, et Tarski nous propose de procéder comme suit :

> Considérons n'importe quel énoncé. Nous le remplacerons par la lettre « p ». Nous formons le nom de cet énoncé et nous le remplaçons par une autre lettre, disons « X ». Nous nous demandons maintenant quelle est la relation logique entre les deux énoncés « X est vrai » et « p ». Il est clair que, du point du vue de notre conception de la vérité, ils sont équivalents. En d'autres termes l'équivalence suivante est valable :
>
> (T) *X est vrai si et seulement si p.*
>
> Nous appellerons chaque équivalence de ce type [...] « une équivalence de la forme (T) [79] ». (Tarski, 1944, p. 272)

Selon Tarski, une fois prise en considération chacune de ces équivalences, qui forment autant de « définitions partielles » de notre conception de la vérité, nous sommes en mesure d'énoncer le critère d'adéquation matérielle :

> Nous sommes maintenant à même de poser de manière précise les conditions sous lesquelles nous considérerons l'usage et la définition du terme « vrai » adéquats du point de vue matériel : nous désirons employer le terme « vrai » de telle manière que toutes les équivalences de la forme (T) puissent être affirmées et *nous appellerons adéquate une définition de la vérité telle que toutes ces équivalences découlent d'elle.* (Tarski, 1944, p. 272-273, italiques de l'auteur)

79. Depuis Tarski, ces biconditionnels sont aussi souvent appelés des **T**-équivalences .

Ce critère d'adéquation matérielle formulé par Tarski est depuis resté particulière-

ment célèbre sous le nom de CONVENTION **T** ou SCHÉMA-**T** [80]. Il a pour but de s'assurer

80. Pour être tout à fait précis d'un point de vue bibliographique, signalons que la CONVENTION **T** dont nous donnons ici une version simplifiée reprise de TARSKI (1944), se trouve formulée pour la première fois dans TARSKI (1935, p. 191), sous une forme à la fois plus précise et un peu plus complexe car donnée dans un cadre plus formalisé :

> Si nous adoptons le symbole 'Vr' pour désigner la classe de tous les énoncés vrais, alors les postulats formulés ci-dessus trouveront leur expression dans la convention suivante :
>
> CONVENTION **T** : La définition du symbole « Vr » formellement correcte, énoncée dans les termes du métalangage, sera appelée définition adéquate de la vérité si elle entraine comme conséquences :
> — (α) tous les énoncés pouvant être obtenus à partir de l'expression
> « $x \in Vr$ si et seulement si p »
> moyennant la substitution au symbole ' x ' du nom décrivant la structure d'un énoncé quelconque formulé dans le langage examiné, et au symbole 'p' de l'expression qui est une traduction en métalangage de l'énoncé donné ;
> — (β) l'énoncé « pour n'importe quel x, si $x \in Vr$ alors $x \in S$ » (ou en d'autres termes '$Vr \subseteq S$' [S désigne ici la classe, précédemment définie dans le texte (p. 182-183), de tous les énoncés du langage-objet])
>
> (TARSKI, 1935, p. 191)

Par rapport à la version simplifiée que nous reprenons dans le corps du texte, plusieurs différences sont à noter : premièrement, Tarski mentionne déjà dans cette formulation la distinction entre langage-objet et métalangage —distinction fondamentale sur laquelle nous allons revenir. Deuxièmement, le nom situé à gauche de la **T**-équivalence n'est pas obtenu par une mise entre guillemets mais constitue une description structurelle de l'énoncé. Tarski parle littéralement de « *structural descriptive names* », *i.e.* de noms des énoncés du langage-objet qui sont construits de façon à refléter leur structure formelle. D'où l'importance d'avoir à faire ici à des langages formalisés, c'est-à-dire à des langages dont la structure syntaxique est parfaitement déterminée et appréhendable. Enfin, et c'est le plus important, l'énoncé situé à droite de l'équivalence et désigné par la lettre 'p' n'est pas l'énoncé du langage-objet lui-même, mais une *traduction* de cet énoncé dans le métalangage. Nous insistons sur ces points car contrairement à ce que pourrait laisser penser la relecture qu'en fera plus tard Quine, la CONVENTION **T** ne se présentait pas au départ sous une forme autorisant à ne voir dans le prédicat de vérité qu'une simple suppression de guillemets de citation. Par ailleurs, remarquons que Tarski reste à peu près silencieux sur la nature de la traduction des énoncés du langage-objet dans le métalangage à l'œuvre ici. Tout au plus laisse-t-il entendre qu'elle doit préserver le contenu de l'énoncé, ou sa « signification intuitive » (?). Pourtant, on pourrait penser qu'il est primordial d'expliquer un peu plus quelles contraintes doit remplir la traduction en question, en particulier si la CONVENTION **T** est censée donner une coniditon d'adéquation extensionnelle pour le prédicat « Vr ».

Si l'on admet pour simplifier (et éviter l'épineux problème de la traduction), que le langage-objet est contenu dans le métalangage, on pourra considérer la traduction dite homophonique où un énoncé est « traduit » par lui-même. Dès lors, si $\ulcorner\urcorner$, désigne un opérateur de citation ou une manière de noter les noms obtenus par description structurelle ou encore un codage gödelien, le critère d'adéquation matérielle de Tarski revient à exiger qu'une théorie de la vérité satisfaisante pour un langage-objet $\mathcal{L}$, formulée dans un métalangage $\mathcal{L}'$ tel que $\mathcal{L} \cup \{Vr\} \subseteq \mathcal{L}'$ *implique* toutes les instances du SCHÉMA-**T** suivantes :

$$\text{(T) } Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$

où φ est un énoncé de $\mathcal{L}$. Ainsi, à gauche de l'équivalence ci-dessus on doit trouver « $\ulcorner\varphi\urcorner$ » un nom/descriptif/code (dans le métalangage $\mathcal{L}'$) de l'énoncé « φ » du langage-objet $\mathcal{L}$, tandis qu'à droite se trouve une traduction homophonique de cet énoncé dans le métalangage, qui est donc identique à l'énoncé lui-même. Chaque instance de ce SCHÉMA-**T** est appelée une **T**-équivalence .

de la correction d'un point de vue extensionnel du prédicat défini, c'est-à-dire de garantir que tomberont dans l'extension du prédicat défini tous les énoncés vrais et uniquement eux. Depuis sa formulation par Tarski, ce critère est généralement considéré comme une exigence d'adéquation extensionnelle minimale que *toute* théorie, définition ou axiomatisation de la vérité doit satisfaire.

Toutefois, nous ne sommes pas encore parvenus à une définition en bonne et due forme. Au cœur de la CONVENTION **T**, on ne trouve pas un énoncé du type

$$\forall x, x \text{ est vrai} \leftrightarrow \phi(x)$$

mettant en relation un *definiendum* avec un *definiens*, mais plutôt un schéma d'énoncés indiquant, pour chaque énoncé fixé, comment construire une équivalence-de-la-forme-**T**. De même, chacune de ces **T**-équivalence prise individuellement, si on peut peut-être la voir comme une « définition partielle », ne constitue visiblement pas une définition générale du prédicat de vérité. Qui plus est, il nous reste également à nous assurer que nos intuitions concernant la conception sémantique de la vérité ne nous ont pas subrepticement menés à des contradictions. Ceci apparaît d'autant plus urgent au vu des antinomies et autres paradoxes sémantiques qui ont émaillé l'histoire des analyses du concept de vérité. Tarski lui-même insiste sur ce point :

> Le problème de la spécification de la structure formelle et du vocabulaire du langage dans lequel doivent être énoncées les définitions des concepts sémantiques devient particulièrement aigu du fait de l'apparition possible des antinomies. (TARSKI, 1944, p. 275)

À la condition d'adéquation matérielle couchée dans CONVENTION **T** vont donc s'ajouter des critères de correction formelle indispensables pour qu'un traitement de la question de la vérité satisfaisant et sûr face au danger des antinomies soit possible. Plus spécifiquement, Tarski met en exergue deux conditions qu'un langage devra remplir. La première est que le langage en question, le langage pour lequel on tente de définir la vérité, possède ce que Tarski appelle une « *structure rigoureusement spécifiée* ». La seconde est qu'il ne soit pas « *sémantiquement clos*[81] ». Commençons par la première :

> *Le problème de la définition la vérité n'acquiert un sens précis et ne peut être résolu d'une manière rigoureuse que pour les langages dont la structure a été rigoureusement spécifiée.* (TARSKI, 1944, p. 276, italiques de l'auteur)

81. Nous allons revenir sur ce que cette expression introduite par Tarski signifie.

Que faut-il entendre par là ? Tout simplement que la grammaire et la syntaxe du langage en question soient parfaitement déterminées et dénuées de toute ambiguïté :

> [nous devons] caractériser de manière non ambiguë la classe de ces mots et expressions qui doivent être considérés comme *doués d'un sens*. (Tarski, 1944, p. 275, italiques de l'auteur)

En d'autres termes, nous devons donc pouvoir caractériser la classe des formes syntaxiques qui possèdent une signification dans le langage, en particulier la classe des termes, des formules et des énoncés. Mais à ce premier réquisit, Tarski en ajoute aussitôt un second : il faudra également que soient formulées

> les conditions sous lesquelles un énoncé peut être asserté. (Tarski, 1944, p. 275)

Nous avons déjà mentionné[82] que, selon un usage courant à son époque, Tarski désigne par un langage non pas seulement un système de signes muni d'une syntaxe, mais également la donnée de conditions d'assertion de certaines expressions du langage en question, c'est-à-dire la donnée de ce que l'on appellerait en termes contemporains une théorie formulée dans le langage. Dans Tarski (1935), il justifie cet emploi de la manière suivante :

> Jusqu'ici on a construit les langages formalisés afin d'élaborer à partir d'eux des *sciences déductives formalisées* ; le langage et la science constituent un tout organique, à tel point qu'au lieu de parler de tel et tel langage formalisé on parle du langage de telle et telle science formalisée. De ce fait, les autres propriétés des langages formalisés [N.D.T. à savoir, la donnée d'axiomes et de règles d'inférence] se manifestent en liaison avec la méthode de constructions des sciences déductives. (Tarski, 1935, § 2, p. 172, ital. de l'auteur)[83]

La spécification de la structure d'un langage consiste donc à caractériser sa syntaxe ainsi que les conditions d'assertion de certains de ces énoncés.

Parmi les langages à structures rigoureusement spécifiées, Tarski distingue plus particulièrement les langages formalisés. Ces langages sont ceux pour lesquels la spécification

82. Voyez 1.1.4.1 page 43.

83. Selon Patterson (2012) cette acception du terme « langage », dans laquelle la donnée d'un langage interprété est liée à la donnée de conditions d'assertabilité, porte également la marque de l'influence sur Tarski des conceptions philosophiques concernant le langage qui avaient cours au sein de l'École de Lvov-Varsovie.

rigoureuse de la structure peut s'effectuer en se référant uniquement à la *forme* des expressions [84]. Autrement dit, dans un langage formalisé,

> le sens de chaque expression est univoquement déterminé par sa forme (Tarski, 1935, p. 172)

C'est pour ce type de langages que Tarski se propose de définir la vérité. De ce fait, seront exclus les langages contenant des indexicaux, des déictiques, des termes vagues ou plus généralement tout langage contenant des expressions ambiguës ou variant avec le contexte. Toutefois, si, comme Tarski lui-même le reconnaît, les seuls langages à structures spécifiées qui ont été actuellement développés sont les langages formels de la logique et des mathématiques [85], Tarski n'entend pas pour autant limiter sa méthode de définition aux seuls langages de ces sciences qu'on a parfois qualifiées de formelles [86]. Bien au contraire, tout langage dont la structure peut être rigoureusement spécifiée est *a priori* susceptible de se voir appliquer la méthode tarskienne, quel que soit son domaine de discours. Dans Tarski (1944, p. 275), Tarski mentionne par exemple le langage de la physique théorique et va même jusqu'à envisager le développement de langages à structures spécifiés non formalisés [87]. Nous voyons donc un peu plus clairement le type de langages pour lesquels Tarski se propose de définir un prédicat de vérité : ce seront des langages à structure rigoureusement spécifiée [88] *et* interprétés, au sein desquels les porteurs de vérité seront les énoncés eux-mêmes considérés comme des types. Se limiter à ce type de langages est apparu nécessaire pour pouvoir donner un sens précis au problème

84. *Cf.* Tarski (1944, p. 275).

85.

> Actuellement, les seuls langages à structure spécifiée sont les langages formalisés des divers systèmes de la logique déductive, éventuellement enrichis par l'introduction de certains termes non logiques. (Tarski, 1944, p. 275)

86. *i.e.* les mathématiques et la logique.

87.

> [...] nous pouvons imaginer la construction de langages qui auraient une structure rigoureusement spécifiée sans être formalisés. Dans un tel langage, l'assertion des énoncés, par exemple, peut ne pas dépendre toujours de leur forme, mais parfois de quelques facteurs non linguistiques. Il serait intéressant et important de construire aujourd'hui un langage de ce type et spécialement un langage dont on pourrait prouver qu'il serait suffisant pour le développement d'une vaste branche de la science empirique. Cela justifierait l'espoir de voir les langages à structure spécifiée remplacer finalement le langage quotidien dans le discours scientifique. (Tarski, 1944, p. 276)

88. De fait, Tarski se limite aux langages formalisés. Les langages à structure spécifiée mais non formalisés, bien qu'ils soient évoqués par Tarski (1944) restent à l'état de projet.

de la définition de la vérité et pour pouvoir espérer se prémunir contre l'apparition des antinomies. L'une des conséquence de cette nécessité d'être muni d'une structure rigoureusement spécifiée est que, pour Tarski, seront exclues du champ d'application de sa méthode les langues naturelles [89].

Mais il y a également une autre raison qui conduit Tarski à douter que la notion de vérité telle qu'elle est employée dans les langues naturelles soit susceptible d'un usage cohérent. Et cette autre raison nous mène à la seconde condition de correction formelle avancée par Tarski. Il s'agit du caractère « *sémantiquement clos* » des langues naturelles. Souvenons-nous en effet que Tarski a énoncé un critère d'adéquation matérielle pour le prédicat de vérité d'un langage sous la forme de la CONVENTION **T**. Il a requis que, pour être adéquate, sa définition soit telle qu'elle permette d'établir toutes les **T**-équivalences pour les énoncés du langage considéré. Or justement, les **T**-équivalences , adoptées sans autre précaution, jouent un rôle crucial dans la dérivation de l'antinomie du Menteur [90]. En voici une illustration semblable à celle de TARSKI (1944, p. 276). Supposons que nous examinions la notion de vérité pour le langage interprété $\mathcal{L}$ = le français [91]. Soit « λ » un nom de l'énoncé imprimé en gras juste à la ligne ci-dessous :

89.

> Pour les autres langages — en premier lieu pour les langages naturels, « parlés »— le sens de ce problème [N.D.T celui de la définition de la vérité] est plus ou moins vague et ses solutions ne peuvent avoir qu'un caractère approximatif. (TARSKI, 1944, p. 276)

Les développements ultérieurs, que ce soient, pour ne citer que les deux exemples les plus célèbres, le programme des grammaires formelles pour les langues naturelles — *cf.* MONTAGUE (1974)— ou le recours à une théorie récursive de la vérité vue comme une théorie de la signification pour répondre au défi de l'interprétation radicale —*cf.* DAVIDSON (1984)— auront (peut-être) montré que l'emploi des méthodes tarskiennes pour l'étude des langues naturelles pouvait s'avérer très fertile. Reste que pour le Tarski des années trente et quarante, les langues naturelles munies de leur prédicat de vérité « pré-théorique » ou « naïf » semblaient irrémédiablement incohérentes (voyez TARSKI (1935, § 1) pour un développement tranchant et sans ambiguïté sur ce point).

90. TARSKI (1944, p. 271, note 4) crédite Lesniewski pour avoir le premier isolé lors de cours non publiés donnés à l'université de Varsovie, l'importance des équivalences de la forme (T) en tant que prémisses de l'antinomie du Menteur.

91. Le lecteur pointilleux pourrait nous faire la remarque suivante : ne sommes-nous pas en train de recourir à une langue naturelle alors même que nous venons précisément de voir que, pour cette classe de langages, le problème de la vérité n'était pas susceptible d'un traitement satisfaisant ? C'est parfaitement exact. Mais, premièrement, en prenant cette liberté, nous suivons les pas de Tarski lui-même (voyez TARSKI (1944)). Deuxièmement, et plus crucialement, il est clair que le raisonnement (*cf.* ci-dessous) sous-tendant la dérivation du Menteur ne repose nullement sur l'absence de spécification rigoureuse du langage en question, mais qu'il peut bien s'appliquer *mutatis mutandis* aux langages à structure spécifiée, sémantiquement clos et obéissant aux lois de la logique classique. Nous renvoyons le lecteur qui demeurerait sceptique à la note 93 où sont données quelques indications sur la façon dont le Menteur peut être dérivé dans le cadre d'un langage parfaitement formalisé.

λ n'est pas vrai.

Alors, sous l'hypothèse que nous puissons donner dans $\mathcal{L}$ une définition du prédicat vrai(-dans-$\mathcal{L}$) adéquate au sens de la CONVENTION **T**, nous devrions être en mesure de dériver la **T**-équivalence correspondant à λ, c'est-à-dire :

λ est vrai si et seulement si λ n'est pas vrai [92]

ce qui implique manifestement une contradiction. La dérivation semble pourtant impeccable. Sur quelles hypothèses nous sommes-nous appuyés pour y parvenir ? TARSKI (1944, p. 277-278) en identifie trois :

1. avoir implicitement admis que le langage $\mathcal{L}$ dans lequel est construite l'antinomie est, selon l'expression de Tarski, un langage « *sémantiquement clos* », par quoi il faut entendre que $\mathcal{L}$ contient non seulement ses propres expressions mais également
   (a) un nom de chacune de ses expressions,
   (b) un prédicat de vérité pour ses énoncés,
   (c) et que l'usage de ce prédicat de vérité est gouverné par les **T**-équivalences , lesquelles sont donc assertables dans ce langage.
2. avoir utilisé les lois de la logique classique,
3. avoir employé une certaine donnée empirique concernant la dénotation du terme « λ », c'est-à-dire avoir constaté que l'énoncé en gras dont λ est un nom n'est autre que l'énoncé « λ n'est pas vrai ».

Tarski remarque rapidement que la prémisse empirique 3. est en réalité inessentielle et qu'on peut reconstruire le paradoxe sans elle [93]. D'autre part, Tarski n'est guère favorable

92. Ou, dans une version plus longue et plus décitationnelle :
   1. « λ n'est pas vrai » est vrai ssi λ n'est pas vrai
   2. λ = « λ n'est pas vrai »
   3. λ est vrai ssi λ n'est pas vrai

93. Les variations sur le paradoxe du Menteur, anciennes ou modernes, sont innombrables. Disons simplement ici que si un langage $\mathcal{L}$ (au sens de Tarski, c'est-à-dire en fait une théorie T exprimée dans $\mathcal{L}$) contient suffisamment d'arithmétique pour coder sa propre syntaxe, alors à supposer que « Vr » soit définissable dans $\mathcal{L}$ et satisfasse la CONVENTION **T**, on obtient facilement l'antinomie en appliquant le lemme de diagonalisation à la formule $\neg Vr(x)$ qui assure de l'existence d'une $\mathcal{L}$-formule λ vérifiant :

$$T \vdash \lambda \leftrightarrow \neg Vr(\ulcorner \lambda \urcorner),$$

à l'idée de renoncer à la logique classique, c'est-à-dire à la prémisse 2. ci-dessus, et il écarte rapidement cette possibilité [94]. Ne reste donc que l'hypothèse 1. ; et Tarski conclut :

> En conséquence, nous prenons la décision de ne pas user d'un langage qui serait sémantiquement clos dans le sens indiqué plus haut. (TARSKI, 1944, p. 278)

Tarski parle ici de décision, mais cette tournure peut paraître assez bizarre, voire trompeuse. Ce que montre l'antinomie du menteur en effet, c'est qu'un langage (classique) et sémantiquement clos est nécessairement inconsistant. Un langage (classique) consistant ne peut donc pas *contenir* un prédicat de vérité pour lui-même adéquat au sens de la CONVENTION **T**. *A fortiori*, il n'est donc pas possible de *définir* adéquatement un prédicat de vérité pour un langage $\mathcal{L}$ dans $\mathcal{L}$ lui-même, si $\mathcal{L}$ est consistant et obéit aux lois usuelles de la logique. Autrement dit, l'antinomie du menteur pose une première limite à la possibilité d'une définition de la vérité dans un cadre logique classique : pour être possible une définition de la vérité pour un langage $\mathcal{L}$ ne pourra pas se faire dans $\mathcal{L}$ lui-même [95]. Il sera donc *nécessaire* de recourir à un autre langage.

En résumé, l'élaboration de critères d'adéquation matérielle et de correction formelle a conduit Tarski à formuler sa fameuse CONVENTION **T** et à restreindre son attention aux langages qui possèdent une structure spécifiée et qui ne sont pas sémantiquement clos. Mais la question demeure : à quelles conditions une définition de la vérité satisfaisant ces contraintes est-elle possible ?

1.1.4.2 Conditions de possibilité : métalangage et « richesse essentielle »

Un aspect fondamental de la solution de ce problème proposée par Tarski va consister à prendre acte de la nécessité de séparer le langage *pour lequel* la vérité est définie et le langage *dans lequel* la définition est conduite. C'est la distinction, aujourd'hui

tandis la **T**-équivalence correspondant à λ nous donne :

$$T \vdash \lambda \leftrightarrow Vr(\ulcorner \lambda \urcorner)$$

94.

> Il serait superflu d'insister ici sur les conséquences du rejet de l'hypothèse (II) [...], c'est-à-dire d'un changement de notre logique (à supposer que cela soit possible) même dans ses parties les plus élémentaires et fondamentales. (TARSKI, 1944, p. 278)

95. Dans sa (ou ses) version(s) formalisée(s), ce célèbre résultat limitatif est depuis resté connu sous le nom de *Théorème d'indéfinissabilité* de Tarski.

classique, entre « langage-objet » et « métalangage ». Le langage-objet est le langage qui constitue l'objet d'étude, en d'autres termes : le langage *pour les énoncés duquel* on cherche à définir la vérité. Le métalangage est le langage *dans lequel* la définition en question, ou plus généralement la « métathéorie[96] » de la vérité, sera développée. La distinction est primordiale car elle permet, sous certaines hypothèses, d'éviter les paradoxes sémantiques ; et l'analyse du Menteur a montré qu'un seul et même langage ne pouvait pas jouer à la fois le rôle de langage-objet et de métalangage, sous peine de contradiction. Les termes du problème sont donc clarifiés et la question à laquelle Tarski s'attaque peut à présent être reformulée comme ceci :

Question. *Soit $\mathcal{L}$ un langage formalisé (et sémantiquement ouvert) ; à quelles conditions peut-on formuler dans un métalangage $\mathcal{M}$ (distinct de $\mathcal{L}$, par nécessité) une métathéorie qui permette de construire une définition du prédicat « vrai-dans-$\mathcal{L}$ » de manière à ce que cette dernière soit formellement correcte et satisfasse la* CONVENTION **T** *?*

La réponse de Tarski, telle qu'il l'exprimera sous une forme non technique dans Tarski (1944), est que la possibilité de formuler une définition de la vérité pour le langage-objet

> [...] dépend de certaines relations formelles existant entre le langage-objet et le métalangage, ou, en termes plus spécifiques, du fait que *le métalangage est ou non, dans sa partie logique, « essentiellement plus riche »*. (Tarski, 1944, p. 281, nous soulignons)

Pour Tarski, cette propriété de « plus grande richesse essentielle » du métalangage se révèle être une condition *nécessaire* pour bloquer l'apparition des antinomies. Mais Tarski va plus loin et affirme que cette condition est également *suffisante* pour garantir la possibilité de donner une définition. C'est ce qu'indique le passage ci-dessous :

> [...] la *condition de « richesse essentielle » du métalangage se révèle être non seulement nécessaire mais encore suffisante pour construire une définition satisfaisante de la vérité*. Autrement dit, si le métalangage remplit cette

96. Nous disons méta*théorie* de la vérité car premièrement, le métalangage dans lequel une définition de la vérité pour le langage-objet est censée être formulée est en fait un langage au sens de Tarski, c'est-à-dire un langage *accompagné d'une théorie* contenant des axiomes et des règles d'inférence. En l'occurrence le métalangage devra au moins contenir une théorie de la morphologie et de la syntaxe du langage-objet. Et, deuxièmement, cette « théorie de la vérité » d'un langage-objet est bien une *méta*théorie par opposition aux théories qui sont formulées dans le langage-objet puisqu'elle s'énonce dans un métalangage et qu'elle a pour objet non pas, ou pas seulement, les choses dont le langage-objet lui même parle, mais plutôt les expressions du langage-objet.

> condition, la notion de vérité peut être définie en lui. (Tarski, 1944, p. 282, ital. de l'auteur)

Tarski semble donc en possession d'une condition à la fois *nécessaire* et *suffisante* caractérisant la possibilité de construction d'une définition de la vérité pour un langage-objet $\mathcal{L}$, à savoir l'existence d'un métalangage $\mathcal{M}$ « essentiellement plus riche ».

Bien entendu, avec une caractérisation de cette nature, nous nous trouvons aussitôt face à une nouvelle interrogation : en quoi consiste exactement cette condition de « plus grande richesse essentielle, dans sa partie logique » d'un métalangage par rapport à un langage-objet ? Malheureusement, comme Tarski lui-même le reconnaît,

> il n'est pas facile de donner une définition générale et précise de cette notion de « richesse essentielle ». (Tarski, 1944, p. 281)

Et si les écrits de Tarski comportent de nombreux indices ou caractérisations partielles (parfois avec une autre terminologie) de ce qu'il voulait désigner par cette expression, on n'y trouve nulle part de caractérisation finale et définitive de cette notion. À tel point que la question de savoir ce que Tarski entendait exactement par « plus grande richesse essentielle » a récemment encore fait l'objet d'un âpre débat exégétique opposant DeVidi et Solomon (1999) et Ray (2005). Sans prétendre proposer de réponse définitive à cette question débattue, nous suivrons Rouilhan (1998) et Ray (2005) pour fournir au lecteur les quelques éléments de compréhension qui suivent.

En s'appuyant sur les sources textuelles de Tarski lui-même (Tarski, 1935, 1933, 1944), il apparaît tout d'abord que la conception tarskienne de plus grande richesse essentielle d'un métalangage par rapport à un langage-objet a évolué au fil du temps dans le sens d'un élargissement de cette notion. La première occurrence de l'expression « essentiellement plus riche » se trouve dans le *Post-Scriptum* de Tarski (1935) :

> Il n'y a évidemment aucun empêchement à introduire des variables d'ordre transfini, non seulement dans le langage qui est l'objet d'investigations, mais encore dans le métalangage dans lequel celles-ci sont menées. En particulier, il est toujours possible de construire le métalangage de telle manière qu'il contienne des variables d'un ordre supérieur aux ordres de toutes les variables du langage étudié. Alors le métalangage devient un langage d'ordre supérieur et partant un langage *essentiellement plus riche* en formes grammaticales que le langage étudié. (Tarski, 1935, p. 263, nous soulignons)

C'est la seule occurence de cette expression dans les différentes versions de la monographie de Tarski (1935, 1933). Dans cet extrait, on constate que la plus grande « richesse essentielle » d'un langage par rapport à un autre concerne les *formes grammaticales* de ces langages et qu'elle est immédiatement reliée à la notion d'*ordre* (des variables) d'un langage. De fait, dans les diverses versions du *Concept de vérité dans les langages formalisés* (Tarski, 1935, 1933), les notions d'ordre supérieur et de plus grande richesse essentielle sont à peu près confondues, et c'est uniquement au moyen de la notion d'ordre que Tarski essaye de caractériser les relations entre, d'une part, les langages-objets pour lesquels on cherche à définir la vérité et, d'autre part, les métalangages censés permettre de donner une définition. Pour comprendre la gestation et l'évolution de la notion de richesse essentielle, il nous faut donc rappeler en quoi consiste la notion d'ordre d'un langage telle qu'elle est développée dans Tarski (1935, 1933).

Dans les premiers temps de ses travaux sur la vérité, c'est-à-dire avant la rédaction du *Post-Scriptum*, Tarski ne prenait en considération que des langages censés obéir à la « théorie des catégories sémantiques » [97]. Comme son nom ne l'indique pas, cette théorie porte principalement sur la structure syntaxique et grammaticale des langages, et les langages qui lui sont soumis sont morphologiquement très proches de ceux de la théorie des types simples. Dans le cadre de cette théorie, chaque variable d'un langage reçoit un ordre fondé sur son rang au sein d'une hiérarchie de catégories sémantiques. Les variables d'individus reçoivent l'ordre 1, les variables de classes d'individus l'ordre 2, les variables de classes de classes d'individus l'ordre 3, *etc.* L'ordre d'un langage est alors déterminé par l'ordre de ses variables. *Grosso modo* l'ordre d'un langage se définit comme le plus petit ordinal supérieur aux ordres de toutes ses variables [98]. Avant le *Postscriptum*, Tarski ne considérait que les langages dont les variables étaient toutes d'ordre fini. Dès lors, un langage se voyait soit attribuer un ordre fini —lorsque les ordres finis de ses variables étaient bornés [99]— soit était simplement déclaré d'ordre infini (sans plus de précision)

97. *Cf.* Tarski (1935) :

> [...] la théorie des catégories sémantiques s'enracine si profondément dans les intuitions fondamentales relatives au sens des expressions, qu'il est impossible d'imaginer un langage scientifique dont les énoncés posséderaient un sens intuitif distinct et dont la structure ne pourrait s'accorder avec cette théorie dans l'une des ses acceptions. (Tarski, 1935, p. 215)

La théorie des catégories sémantiques a été originellement introduite par Husserl avant d'être reprise par certains membres de l'École de Lvov-Varsovie, notamment Lesniewski. Pour plus de renseignements sur l'origine de la théorie des catégories sémantiques chez Tarski voyez Rouilhan (1998).

98. Nous simplifions largement ici. Pour plus de précisions sur la notion d'ordre dans Tarski (1935, 1933) nous renvoyons aux textes de Tarski lui-même, ainsi qu'à Rouilhan (1998).

99. l'ordre du langage était alors le sup des ordres de ses variables.

—lorsque les ordres (finis) de ses variables étaient non bornés [100].

Avant la rédaction du *Post-Scriptum* et avec la notion d'ordre que nous venons de rappeler, Tarski (1933) aboutissait alors au double résultat suivant concernant la possibilité de définir la vérité :

A. un théorème positif [101] : pour les langages d'ordre *fini*, une définition adéquate de la vérité est possible en se plaçant dans un métalangage d'ordre supérieur ;

B. et un théorème négatif d'impossibilité [102] : dans le cas des langages d'ordre *infini*, il n'est pas possible de les enchâsser dans un métalangage d'ordre supérieur et leurs prédicats de vérité sont donc « absolument » indéfinissables [103].

La situation change radicalement avec le *Post-Scriptum*, et les conceptions de Tarski connaissent une inflexion fondamentale. Dans ce texte ajouté à l'occasion de la parution de la traduction allemande, Tarski renonce purement et simplement à la théorie des catégories sémantiques. Alors que précédemment il considérait que tout langage scientifique doté d'une signification devait pouvoir s'accorder avec ce cadre, dorénavant il entend au contraire accepter et étudier

> l'introduction dans le champ de nos investigations des langages formalisés pour lesquels les principes de la théorie des catégories sémantiques ne seraient plus valables. (Tarski, 1935, p. 260)

L'une des conséquences de ce changement de point de vue est qu'à présent Tarski est prêt à envisager l'existence d'ordres infinis non plus seulement pour les langages eux-mêmes mais également pour certaines expressions (notamment des variables) contenues dans

100. Par exemple, lorsqu'un langage $\mathcal{L}$ contenait des variables d'ordre fini n pour tout $n \in \mathbb{N}$. Chaque variable est d'ordre fini, mais l'ordre de $\mathcal{L}$ lui-même est supérieur à tout n, donc infini.

101. C'est la thèse A. de la conclusion pré-*Post-Scriptum* de Tarski (1935) :

> A. Nous savons construire dans le métalangage, pour chaque langage formalisé d'ordre fini, une définition formellement correcte et matériellement adéquate de la notion d'énoncé vrai [...] (Tarski, 1935, p. 258)

102. C'est la thèse B. de la conclusion pré-*Post-Scriptum* de Tarski (1935) ou d' « indéfinissabilité absolue » :

> B. Il n'est pas possible de construire une telle définition pour les langages formalisés d'ordre infini. (Tarski, 1935, p. 258)

103. Nous disons *absolument* indéfinissable pour signifier que, étant donné un langage $\mathcal{L}$ d'ordre infini, la vérité-pour-$\mathcal{L}$ est non seulement indéfinissable dans $\mathcal{L}$ lui-même, mais encore indéfinissable « tout court », puisqu'il n'existe pas de métalangage $\mathcal{M}$ d'ordre supérieur à celui de $\mathcal{L}$, ou pour le dire autrement et en anticipant la terminologie que Tarski adoptera par la suite, il n'existe pas de métalangage $\mathcal{M}$ essentiellement plus riche que $\mathcal{L}$.

ces langages :

> [...] il faut tenir compte du fait qu'il est possible que d'autres foncteurs permettant de former des énoncés surviennent dans le langage donné auxquels un ordre infini doit être assigné. Si, par exemple, un signe n'est un foncteur permettant de former des énoncés que dans des formules dont tous les arguments sont d'ordre fini, mais où leurs ordres ne sont cependant pas limités d'avance par un nombre naturel, alors *ce signe est d'ordre infini*. (Tarski, 1935, p. 261, nous soulignons)

Autre innovation importante accompagnant cette évolution : désormais pour Tarski, les ordres *infinis* des expressions ou des langages sont eux- mêmes susceptibles d'être classés entre eux. Il suffit pour ce faire de s'appuyer sur la notion d'ordinal issue de la théorie des ensembles :

> Afin de classer les signes d'ordre infini, nous recourons à la notion de *nombre ordinal* empruntée à la théorie des ensembles. [...] En fait, nous pouvons assigner à chaque langage un nombre ordinal bien spécifique indiquant son ordre, à savoir le plus petit ordinal dépassant les nombres [104] des ordres de toutes les variables figurant dans ce langage. (Tarski, 1935, p. 261-262)

Autrement dit, la hiérarchie classant les ordres des langages qui, dans le corps du texte précédant le *Post-Scriptum* se terminait à l'infini, est désormais prolongée au transfini. Comme Tarski lui-même le souligne,

> [...] de ce fait disparaît la différence entre les langages d'ordre fini et d'ordre infini, différence qui a été si importante dans les § § 4 et 5 et qui a trouvé une forte expression dans les thèses A et B de la conclusion [105] (Tarski, 1935, p. 263)

Il est en effet bien connu que la classe des ordinaux ne possède pas de borne supérieure au sens où pour tout ordinal (fût-il infini) fixé, il existe toujours un ordinal qui lui est strictement supérieur. Dès lors, puisque tout langage se voit attribuer un ordinal pour mesure de son ordre, rien n'empêche d'imaginer, pour tout langage-objet d'ordre fixé

104. [N.D.T. : ici « nombre » est à prendre au sens de nombre ordinal, c'est-à-dire q'il peut s'agir de « nombres transfinis ».]

105. Tarski fait ici référence au double résultat obtenu dans le corps du texte et que nous avons déjà rappelé. Voyez ci-dessus page 56.

même infini, l'existence d'un métalangage d'ordre strictement supérieur. C'est exactement ce qui est au cœur du passage que nous avons déjà cité sur la notion de richesse essentielle :

> [...] il est *toujours possible* de construire le métalangage de telle manière qu'il contienne des variables d'un ordre supérieur aux ordres de toutes les variables du langage étudié. Alors le métalangage devient *un langage d'ordre supérieur* et partant un langage essentiellement plus riche [...] (Tarski, 1935, p. 263, nous soulignons)

Avec cette modification de la notion d'ordre,

> [...] la construction d'une définition correcte de la vérité pour les langages d'ordre infini [devient] en principe possible pourvu que nous ayons à notre disposition, dans le métalangage, des expressions d'un ordre supérieur à celui de toutes les variables du langage étudié. [...] maintenant nous sommes en mesure de définir le concept de vérité pour tout langage d'ordre fini ou infini pourvu que nous prenions pour base de nos investigations le métalangage d'un ordre plus grand d'une unité au moins que le langage étudié [...]. (Tarski, 1935, p. 263)

Ce renversement aboutit au remplacement du double résultat obtenu dans la conclusion pré-*Post-Scriptum* (y compris le résultat d'indéfinissabilité absolue) [106] par les deux nouvelles thèses modifiées suivantes :

> A. Une définition correcte et matériellement adéquate de l'énoncé vrai peut être construite pour tout langage formalisé [...] à condition [...] que le métalangage possède un ordre supérieur à celui du langage étudié.
>
> B. La construction d'une telle définition n'est pas possible si l'ordre du métalangage est tout au plus égal à celui du langage étudié.
>
> (Tarski, 1935, p. 264)

En d'autres termes, il n'y a plus, pour le Tarski du *Post-Scriptum*, de langages-objets dont la vérité soit absolument indéfinissable ; seule demeure la nécessité pour pouvoir construire une telle définition de se placer dans un métalangage d'ordre supérieur, ce qui, aux yeux de Tarski, est dorénavant toujours possible [107].

106. Voyez page 56.
107. Selon Rouilhan (1998), ce changement des conceptions de Tarski constitue un véritable change-

Mais cette modification, déjà capitale, n'est pas la seule opérée sur la notion d'ordre à l'occasion du *Post-Scriptum*. Une autre révision substantielle est évoquée dans une note de bas de page de la plus haute importance (*cf.* Tarski (1935, p. 262-263, note 3)). En renonçant à la théorie des catégories sémantiques, Tarski s'ouvre la possibilité d'examiner des

> langages d'un autre genre qui constituent un outil de développement de la logique et des mathématiques beaucoup plus convenable et actuellement beaucoup plus souvent employé. (Tarski, 1935, p. 262-263, note 3)

Ces langages sont ceux de « Zermelo et ses successeurs ». Crucialement,

> [d]ans ces nouveaux langages toutes les variables sont d'un *ordre indéterminé*. (Tarski, 1935, p. 262-263, note 3, nous soulignons)

Tarski caractérise d'ailleurs ces variables d'ordre indéterminé en les qualifiant de

> [...] variables qui « parcourent » pour ainsi dire tous les ordres possibles [...] (Tarski, 1935, p. 262)

Dans le langage de la théorie des ensembles tel qu'il a été développé par Zermelo et ses successeurs et tel que nous le connaissons aujourd'hui, les variables sont effectivement toutes placées sur un même pied d'égalité d'un point de vue syntaxique. Et dans ces langages, une même variable peut prendre pour valeur un individu, un ensemble d'individus, un ensemble d'ensembles d'individus, un ensemble de tels ensembles, *etc.* Les contraintes syntaxiques de la théorie des catégories sémantiques, semblables à celles des langages de la théorie des types simples, où chaque variable renvoie à une catégorie précise d'objets n'ont plus cours ici. Pour autant, selon Tarski,

> [l]e concept d'ordre ne perd d'aucune manière son importance pour les langages étudiés ici [N.D.T. *i.e.* les langages aux variables d'ordre indéterminé de Zermelo et ses successeurs]. Cependant *il ne s'applique plus aux expressions du langage donné*, mais *soit aux objets désignés par ces expressions, soit au langage tout entier*. Nous appelons les objets d'ordre 0 individus, c'est-à-dire les objets qui ne sont pas des classes ; l'ordre d'une classe quelconque

ment de paradigme et Tarski a tort de le présenter sous le faux jour de l'évidence mathématique. En réalité, d'après Rouilhan (1998) le prix à payer pour dépasser l'indéfinissabilité absolue de la vérité pour les langages d'ordre infini est le renoncement à la possibilité d'un langage authentiquement universel et permettant de parler en toute généralité, autrement dit l'abandon de l'universalisme logique classique au profit d'une conception plus « modèle-théorique » de la logique. Nous renvoyons au texte de Rouilhan (1998) pour une discussion argumentée de cette question.

> correspond au plus petit nombre ordinal supérieur aux ordres de tous les éléments de la classe donnée ; l'ordre du langage est représenté par le plus petit nombre ordinal dépassant l'ordre de *toutes les classes dont l'existence découle des axiomes adoptés dans le langage*. (Tarski, 1935, p. 263, note 3, nous soulignons)

On détecte ici une autre transformation essentielle de la notion d'ordre d'un langage : alors que précédemment l'ordre d'un langage était caractérisé comme « le plus petit ordinal dépassant les nombres [ordinaux] des ordres de toutes les variables figurant dans ce langage », dans cet extrait, lorsqu'on examine le cas d'un langage dont les variables sont elles-mêmes d'ordre indéterminé, on constate que la notion d'ordre « ne s'applique plus aux expressions du langage donné » mais est défini comme le plus petit ordinal supérieur à « l'ordre de *toutes les classes dont l'existence découle des axiomes adoptés dans le langage* ». On passe donc d'une conception « syntaxique » de la notion d'ordre, héritée de la théorie des catégories sémantiques quoiqu'éventuellement généralisée pour prendre en compte les expressions d'ordre infini, à une conception qu'on est tenté de qualifier d'« ontologique » et qui s'appuie non plus sur les formes syntaxiques ou grammaticales contenues dans le langage mais sur les objets (plus précisément les classes) dont l'existence peut être établie à partir des axiomes du langage [108]. Cette nouvelle notion d'ordre renvoie *in fine* à la place occupée au sein de la hiérarchie cumulative des ensembles, par les classes (ou plus précisément les ensembles) dont l'existence découle des axiomes du langage [109]. Et, en ce sens nouveau, un métalangage sera d'ordre supérieur à un langage-objet dès lors qu'on pourra établir à partir des axiomes de ce métalangage l'existence d'une classe située plus haut dans la hiérarchie cumulative que toutes les classes dont l'existence découle des axiomes du langage-objet. Il y a évidemment là, semble-t-il, un changement majeur de conception.

En résumé, avec l'ajout du *Post-Scriptum* dans Tarski (1935), la notion d'ordre —et celle de richesse essentielle avec laquelle elle se confond dans ce texte— connaît une double évolution : tout d'abord, à l'aide de la théorie des ordinaux, la hiérarchie des ordres est prolongée au transfini ; mais d'autre part, pour traiter le cas des langages aux variables d'ordre indéterminé, c'est-à-dire dont les variables « parcourent pour ainsi dire

108. Puisque, rappelons-le une fois encore, quand Tarski parle de langage, il désigne en fait toujours le langage d'une théorie, c'est-à-dire un langage muni d'axiomes et de règles d'inférences.
109. Sur ce point précis, sur lequel nous ne nous étendrons pas davantage ici, voyez Ray (2005) et surtout Rouilhan (1998).

tous les ordres possibles », une nouvelle notion d'ordre est introduite, qui renvoie à la possibilité de démontrer l'existence de certaines classes à partir des axiomes du langage considéré.

Toutefois, l'évolution de la notion de « plus grande richesse essentielle » ne s'arrête pas là. Et la situation se complique encore par la suite puisque Tarski revient de nouveau sur cette notion pour y apporter une nouvelle inflexion. Dans une note de bas de page tirée de Tarski (1944), Tarski introduit une nouvelle manière pour un langage d'être « essentiellement plus riche » qu'un autre. Voici l'extrait en question :

> Cette esquisse en gros ne permet pas de voir à quel endroit et de quelle manière la supposition de la « richesse essentielle » du métalangage est insérée dans la discussion. Cela ne devient clair que lorsque la définition de la vérité est exposée de manière formelle et dans tous ses détails. [*Note de Tarski : Afin de définir récursivement la notion de satisfaction, nous devons appliquer une certaine forme de définition par récurrence laquelle n'est pas admise dans le langage-objet. C'est pourquoi la « richesse essentielle » du métalangage peut consister tout simplement à admettre ce genre de définition.* D'autre part, on connaît la méthode générale permettant d'éliminer toutes les définitions récursives et de les remplacer par des définitions normales, explicites. Si nous essayons d'appliquer cette méthode à la définition de la satisfaction, nous voyons que nous avons, soit à introduire dans le métalangage des variables d'un langage d'un type logique supérieur à celles qui appartiennent au langage-objet, soit à admettre par voie d'axiomes l'existence de classes plus riches que toutes celles dont l'existence peut être établie en langage-objet.] (Tarski, 1944, p. 284, corps de texte et note 13, nous soulignons)

Ici Tarski se concentre sur le type de définition acceptable au sein d'un langage. La différence de richesse essentielle peut consister pour un métalangage $\mathcal{M}$ à être simplement muni de formes de définitions acceptables supplémentaires par rapport à un langage-objet $\mathcal{L}$ [110]. À bien y regarder d'ailleurs, même dans Tarski (1935, 1933), le rôle joué par les

110. Nous laisserons de côté la question de savoir ce que peut signifier exactement pour un langage d'« accepter » tel ou tel type de définition. Nous nous contenterons de renvoyer une fois de plus au fait que, sous la plume de Tarski, le mot « langage » désigne en fait non seulement un ensemble de signes muni d'une syntaxe (c'est-à-dire un langage au sens moderne et strict du terme), mais également une théorie exprimée dans ce langage, c'est-à-dire un système d'axiomes muni de règles d'inférence et, visiblement, de règles de définition.

variables d'ordre supérieur du métalangage consistait essentiellement à rendre possible la transformation d'une définition récursive en une définition explicite en bonne et due forme, selon une technique classique de logique bien connue depuis Frege. Si on relâche l'exigence d'obtention d'une définition explicite [111], d'autres possibilités apparaissent, notamment celle d'employer des définitions récursives ou bien plus généralement de se contenter d'une axiomatisation du prédicat de vérité introduit comme un terme primitif non défini [112].

Au total, selon Ray (2005, p. 437) on obtient donc une notion disjonctive de « richesse essentielle » que l'on peut caractériser comme suit :

> Dire qu'un métalangage $\mathcal{M}$ *est essentiellement plus riche qu'*un langage-objet $\mathcal{L}$, c'est dire qu'une des conditions suivantes est remplie (i) $\mathcal{M}$ est un langage fondé sur la théorie des catégories sémantiques et possède des variables d'ordre supérieur à $\mathcal{L}$, (ii) les axiomes de $\mathcal{M}$ démontrent l'existence de quelque chose d'ordre supérieur à tout ce dont on peut prouver l'existence à partir des axiomes de $\mathcal{L}$, ou (iii) les règles de définition de $\mathcal{M}$ admettent les définitions récursives tandis que les règles de définition de $\mathcal{L}$ les rejettent. (Ray, 2005, p. 437, ital. de l'auteur)

Il apparaît ainsi que la condition de plus grande richesse essentielle d'un métalangage par rapport à un langage-objet consiste pour ce métalangage à être pourvu de moyens expressifs plus riches, que ce soit par la présence de formes grammaticales supplémentaires, par la position d'axiomes plus forts ou par l'acceptation de formes nouvelles de définitions, ces moyens expressifs plus riches devant rendre possible une définition de la vérité pour le langage-objet tout en se gardant des paradoxes. Dans le cadre de ce travail, tel sera notre dernier mot sur ce point. Et nous ne tenterons pas de caractériser plus précisément la notion tarskienne de plus grande « richesse essentielle », qui demeure une question complexe et controversée [113].

Sur les caractéristiques techniques des définitions récursives, nous renvoyons à la littérature logico-mathématique consacrée à cette question. Le classique Moschovakis (1974) est un bon point d'entrée.

111. et qu'on renonce aux garanties de cohérence qu'elle semble fournir et qui étaient si centrales pour le Tarski des années 1930, mais l'étaient peut-être déjà moins pour le Tarski des années 1940.

112. Nous reviendrons sur ce point par la suite.

113. Pour plus de détails, nous renvoyons à nouveau à DeVidi et Solomon (1999) et Ray (2005). Voyez également Rouilhan (1998), en particulier sur la notion tarskienne de (méta)langage d'ordre supérieur (à un langage-objet), et sur son évolution entre le corps du texte de Tarski (1933) et le *Post-Scriptum* ajouté à l'occasion l'édition allemande Tarski (1935). Signalons également que, selon Patterson (2012), la solution du mystère de cette notion de « richesse essentielle » nous est donnée par Tarski lui-même dans un autre texte moins fréquemment mentionné. Voici le passage de Tarski en

Ceci étant, il vaut en revanche la peine d'examiner plus en détails la façon dont Tarski justifie la nature nécessaire et suffisante de cette plus grande richesse essentielle comme condition de possibilité d'une définition de la vérité. La démonstration du caractère nécessaire de cette condition se fait essentiellement par la formalisation de l'antinomie du Menteur et nous ne nous y attarderons pas. Dans Tarski (1944), Tarski donne les indications suivantes sur ce point :

> Si la condition de « richesse essentielle » n'est pas remplie, on peut habituellement montrer qu'une interprétation du métalangage dans le langage-objet est possible. Cela veut dire qu'avec un terme donné du métalangage peut être mis en relation un terme bien défini du langage-objet, de telle manière que les énoncés susceptibles d'assertion dans un langage se trouvent correspondre aux énoncés susceptibles d'assertion dans l'autre. En conséquence, l'hypothèse selon laquelle une définition satisfaisante de la vérité est formulée dans le métalangage se trouve impliquer la possibilité d'une reconstruction dans ce langage de l'antinomie du menteur. Et cela nous oblige en retour à rejeter l'hypothèse en question. (Tarski, 1944, p.)

Autrement dit, si la condition de plus grande richesse essentielle n'est pas satisfaite, on peut réinterpréter le métalangage $\mathcal{M}$ à l'intérieur du langage-objet $\mathcal{L}$ lui-même, ce qui, sous l'hypothèse que $\mathcal{M}$ contient une définition adéquate de la vérité-pour-$\mathcal{L}$, fait de $\mathcal{L}$ un langage sémantiquement clos dans lequel on pourra donc reconstituer le raisonnement menant à l'antinomie du Menteur. Ceci établit le caractère nécessaire de la condition de plus grande richesse essentielle [114].

question, cité *in* (Patterson, 2012, p. 70-71) :

> [...] nous introduisons d'abord un concept auxiliaire. Soit X et Y deux ensembles d'énoncés. Nous dirons que l'ensemble Y est *essentiellement plus riche que l'ensemble X par rapport aux termes spécifique*, si : (1) tout énoncé de l'ensemble X appartient à l'ensemble Y (et donc tout terme spécifique de X se trouve également dans l'ensemble Y), et si : (2) on trouve dans les énoncés de Y des termes spécifiques qui ne sont pas dans les énoncés de X et qui ne peuvent être définis, même sur la base de l'ensemble Y, uniquement à l'aide des termes de l'ensemble X. (Tarski, 1956b, p. 37, ital. de l'auteur, la pagination renvoie à la traduction française),

Patterson (2012) fait suivre cette citation du commentaire suivant :

> Une certaine quantité d'encre, pour une bonne part critique (e. g. (DeVidi et Solomon, 1999)), a été versée sur cette question de savoir ce que Tarski voulait dire par un langage « essentiellement plus riche », mais cela nous est exactement expliqué ici. Voyez [...] Ray (2005) pour une réponse à DeVidi et Solomon. (Patterson, 2012, note 14 p. 238).

114. Pour un exposé plus détaillé sur un plan technique, nous renvoyons à Tarski (1935), en particulier

La manière dont Tarski justifie le caractère suffisant de la condition de richesse essentielle s'est, quant à elle, avérée encore plus fertile et intéressante pour notre propos. Nous allons donc nous y attarder un peu plus. La stratégie employée par Tarski ne consiste pas à donner de ce résultat une preuve directe en toute généralité [115]. Au contraire, Tarski fixe un langage (et une théorie) objet donné(s) [116] et il se place ensuite dans un métalangage essentiellement plus riche à l'intérieur duquel il montre comment on peut construire une définition de la vérité pour le langage-objet, qui satisfera tous les critères et conditions qu'il a précédemment énoncés. La construction exposée est assez « archétypique » pour pouvoir être généralisée [117]. Nous suivrons ici l'exemple de Tarski et nous allons donner un exemple de construction d'une définition de la vérité pour un langage-objet donné dans un métalangage plus riche. Ce sera l'occasion d'exposer la variété des techniques tarskiennes [118].

1.1.4.3 La construction d'une définition

Commençons par quelques remarques préliminaires. Soit $\mathcal{L}$ un langage-objet formalisé. Comme le rappelle Tarski,

> Le vocabulaire du métalangage est pour une large part déterminée par les conditions établies précédemment sous lesquelles une définition de la vérité peut être considérée comme matériellement adéquate. (Tarski, 1944, p.)

à la démonstration du théorème I (Tarski, 1935, §5, p. 243).

115. Ce qui d'ailleurs semble assez difficile puisqu'il faudrait pour cela considérer toutes les paires possibles de langage-objets/métalangages $\langle \mathcal{L}, \mathcal{M} \rangle$ avec $\mathcal{M}$ essentiellement plus riche que $\mathcal{L}$ et montrer qu'une définition de vrai-pour-$\mathcal{L}$ est toujours possible dans $\mathcal{M}$.

116. Historiquement, dans le cas précis de Tarski (1935), Tarski choisit le langage (et la théorie) du calcul des classes (*cf.* Tarski (1935, § 2 & 3, p. 172-209)).

117. au moins à une large classe de langages formalisés. En l'occurrence, selon Tarski, à tous les langages formalisés qui ont été développés jusqu'à présent :

> La méthode [...] peut être appliquée —avec des changements appropriés— à tous les langages formalisés connus jusqu'ici, bien qu'il ne s'ensuive pas qu'il serait impossible de construire un langage auquel elle ne s'appliquerait pas. (Tarski, 1944, p. 282, note 11)

Voyez aussi les sections § 4 et 5 de Tarski (1935), où Tarski explique comment généraliser sa méthode :

> La méthode de construction utilisée dans le paragraphe précédent pour l'étude du langage de l'algèbre des classes se laisse appliquer, avec des modifications qui ne sont pas tellement essentielles, à plusieurs autres langages formalisés, même si leur structure logique est considérablement plus compliquée. Les remarques qui suivent ont pour but de mettre en relief le caractère général de cette méthode, de tracer les limites de son application et d'esquisser les modification auxquelles elle reste soumise dans divers cas concrets d'application. (Tarski, 1935, p. 210, début de la section § 4)

118. Ou plutôt d'en donner un aperçu raisonnablement non technique.

Ainsi, lorsque pour construire une définition de la vérité-pour-$\mathcal{L}$ formellement correcte et satisfaisant la CONVENTION **T**, nous nous plaçons dans un métalangage $\mathcal{M}$ essentiellement plus riche, ce dernier devra contenir à tout le moins —ne serait-ce que pour pouvoir formuler les **T**-équivalences — les éléments suivants :

1. pour chaque énoncé du langage-objet $\mathcal{L}$, un énoncé ayant la même signification. Autrement dit, une traduction dans $\mathcal{M}$ de chaque énoncé de $\mathcal{L}$;
2. les ressources nécessaires pour pouvoir construire un nom, ou une description structurelle, de chaque énoncé du langage-objet ;
3. des termes de caractère logique et ensembliste général [119], c'est-à-dire des expressions telles que « si et seulement si » mais également les ressources nécessaire pour pouvoir formuler les notions indispensables à la définition de la satisfaction (théorie des suites d'objets, etc) [120].

Voilà pour les ressources lexicales du métalangage. Mais rappelons-nous une fois encore que par « langage », Tarski désigne aussi la donnée d'axiomes et de règles d'assertion. Notre métalangage contiendra donc aussi une métathéorie suffisante pour conduire la définition de la vérité, c'est-à-dire des axiomes de trois types :

1. des axiomes ayant la même signification que ceux du langage-objet [121] ;
2. des axiomes permettant de développer une théorie de la syntaxe et de la morphologie de $\mathcal{L}$;
3. des axiomes logiques (et ensemblistes) généraux.

À présent, supposons que les conditions ci-dessus soient réunies et voyons plus en détails comment fonctionne une définition tarskienne de la vérité. Comme TARSKI (1935, p. 192) lui-même l'indique, si notre langage-objet $\mathcal{L}$ ne contenait qu'un nombre fini d'énoncés, disons $p_1, \ldots, p_n$, et que notre métalangage $\mathcal{M}$ contenait à la fois une *traduction* $\phi_1, \ldots, \phi_n$ [122] ainsi qu'un *nom* $x_1, \ldots, x_n$ pour chacun d'entre eux, alors il y aurait un moyen très simple de définir dans $\mathcal{M}$ un prédicat de vérité « Vr » pour $\mathcal{L}$, adéquat au sens de la CONVENTION **T**. Il suffirait pour cela de recourir à une élémentaire définition stipulative de la forme :

119. *Cf.* TARSKI (1944, p. 280)
120. Voir plus bas.
121. Là encore, nous parlons de langage-objet au sens de Tarski, c'est-à-dire en termes contemporains, d'un langage au sens strict, accompagné d'une théorie formulée dans ce langage. C'est en référence à cette dernière que nous parlons des axiomes du langage-objet. Notez toutefois que la théorie-objet peut éventuellement être la théorie vide.
122. Autrement dit, un énoncé ϕ_i de $\mathcal{M}$ ayant la même signification que l'énoncé p_i de $\mathcal{L}$ correspondant.

$$\text{(Def) } Vr(x) \leftrightarrow ((x = x_1 \text{ et } \phi_1) \text{ ou } (x = x_2 \text{ et } \phi_2) \text{ ou } \ldots \text{ ou } (x = x_n \text{ et } \phi_n))$$

On peut facilement prouver qu'une telle définition obéit à la CONVENTION **T**, autrement dit en dériver chacune des n **T**-équivalences suivantes ($1 \leq i \leq n$) :

$$Vr(x_i) \leftrightarrow \phi_i$$

Néanmoins, il est évident qu'une définition de ce genre est impossible pour un langage contenant un nombre infini d'énoncés (comme c'est le cas, par exemple, pour le langage de la logique propositionnelle). C'est ici que l'hypothèse des langages à « structure rigoureusement spécifiée » intervient de manière cruciale. En effet, Tarski remarque que si la syntaxe du langage-objet $\mathcal{L}$ est rigoureusement spécifiée et que la signification [123] d'un énoncé complexe *ne* dépend *que* de la signification de ses constituants et de la manière dont ils sont combinés, alors il sera parfois possible de s'appuyer sur cette structure pour donner une définition *récursive* de la vérité.

En voici une première illustration particulièrement simple : supposons que notre langage-objet $\mathcal{L}$ soit construit à partir d'un nombre fini d'énoncés « élémentaires [124] » $p_1, \ldots, p_n$, au moyen d'un unique opérateur vériconditionnel « $\wedge$ », de telle sorte que tout $\mathcal{L}$-énoncé complexe « φ » = « $\varphi_1 \wedge \varphi_2$ », construit de manière univoque à partir de deux énoncés plus simples « φ_1 » et « φ_2 » au moyen de l'opérateur « $\wedge$ », voit sa valeur de vérité *ne* dépendre *que* de celles de ses constituants et de l'opérateur employé pour les combiner. Par exemple, si « $\wedge$ » est un opérateur de conjonction, alors « φ » sera vrai si et seulement si « φ_1 » et « φ_2 » sont tous les deux vrais. Supposons également que note métalangage $\mathcal{M}$ contiennent les ressources pour construire un nom de chacun des énoncés ψ (en nombre infini) de $\mathcal{L}$, que nous noterons $\ulcorner\psi\urcorner$. Supposons en outre que $\mathcal{M}$ contienne des ressources logiques suffisantes, notamment une conjonction « et », et une disjonction « ou ». Supposons enfin que $\mathcal{M}$ comprenne un traduction de chaque énoncé

123. ou à tout le moins la valeur de vérité.
124. Ou atomiques, selon la terminologie courante de la logique propositionnelle.

ψ de $\mathcal{L}$, que nous noterons ψ^*. Alors, on peut construire la définition récursive suivante :

$$
\begin{aligned}
(Def)\ Vr(x) \leftrightarrow \Big(& (x = \ulcorner p_1 \urcorner \text{ et } p_1^*) \text{ ou} \\
& (x = \ulcorner p_2 \urcorner \text{ et } p_2^*) \text{ ou} \\
& \ldots \\
& (x = \ulcorner p_n \urcorner \text{ et } p_n^*) \text{ ou} \\
& \Big(x = \ulcorner \varphi_1 \wedge \varphi_2 \urcorner \text{ et } (Vr(\ulcorner \varphi_1 \urcorner) \text{ et } Vr(\ulcorner \varphi_2 \urcorner))\Big)\Big)
\end{aligned}
$$ [125]

Là encore, il est assez aisé de montrer (par induction sur la complexité des formules de $\mathcal{L}$) que le prédicat ainsi défini satisfait la CONVENTION **T** [126].

Voilà donc le genre de chemin que Tarski propose d'emprunter pour parvenir à une définition de la vérité. Toutefois, nous ne sommes pas encore au bout de nos peines. Dans le cas des langages quantifiés, un autre obstacle se trouve en travers de la route et Tarski va trouver une solution très ingénieuse pour le franchir. L'obstacle en question réside dans le fait que certains énoncés complexes des langages quantifiés sont obtenus

125. Deux points sur lesquels nous voudrions attirer l'attention du lecteur :

1. Remarquez que si nous reprenions les notations de l'exemple précédent, les n premiers membres de la disjonction composant la définition s'écriraient de manière rigoureusement identique :

$$(x = x_1 \text{ et } \phi_1) \text{ ou } (x = x_2 \text{ et } \phi_2) \text{ ou } \ldots \text{ ou } (x = x_n \text{ et } \phi_n).$$

On ne fait qu'ajouter une $(n+1)^{\text{e}}$ clause à la disjonction pour traiter le cas des énoncés composés.

2. Telle qu'elle, la définition n'est pas une définition (explicite) en bonne et due forme puisque le terme défini « Vr » apparaît lui-même dans le *definiens*, *i.e.* à droite du « $\leftrightarrow$ ». Néanmoins, il existe des techniques classiques permettant de transformer les définitions récursives de ce genre en définitions explicites. Pour cela, on a généralement besoin d'employer des variables d'ordre supérieur à celles qui apparaissent dans la définition. Ici, le fait de se placer dans un métalangage $\mathcal{M}$ essentiellement plus riche que $\mathcal{L}$ peut donc avoir une importance pour obtenir une définition explicite en bonne et due forme (voyez aussi RAY (2005) et les rappels que nous avons faits précédemment).

126. En guise d'illustration, voici l'ébauche d'une dérivation d'une **T**-équivalence pour l'énoncé $\ulcorner p_1 \wedge p_2 \urcorner$: la dernière clause de la définition de « Vr » nous donne rapidement $Vr(\ulcorner p_1 \wedge p_2 \urcorner) \leftrightarrow (Vr(\ulcorner p_1 \urcorner) \text{ et } Vr(\ulcorner p_2 \urcorner))$. Pour $\ulcorner p_1 \urcorner$ et $\ulcorner p_2 \urcorner$, les clauses correspondantes de la définition permettent d'obtenir $Vr(\ulcorner p_1 \urcorner) \leftrightarrow p_1^*$ ainsi que $Vr(\ulcorner p_2 \urcorner) \leftrightarrow p_2^*$. Après quelques manipulations logiques et en s'appuyant sur la théorie de la syntaxe de $\mathcal{L}$ contenue dans $\mathcal{M}$, on obtient d'abord

$$Vr(\ulcorner p_1 \wedge p_2 \urcorner) \leftrightarrow p_1^* \text{ et } p_2^*,$$

puis

$$Vr(\ulcorner p_1 \wedge p_2 \urcorner) \leftrightarrow (p_1 \wedge p_2)^*$$

en composant des constituants qui, eux-mêmes, ne sont pas des énoncés susceptibles de prendre une valeur de vérité mais des formules ouvertes. La solution découverte par Tarski consiste à faire un détour par une notion intermédiaire qu'il introduit, à savoir la notion de satisfaction, avant de définir la vérité à proprement parler. Pour exposer plus précisément comment cela fonctionne, il nous faut entrer encore un peu plus dans les détails techniques et les joies de la syntaxe et de la morphologie. À la suite de Tarski (1935), nous examinerons donc un peu plus en profondeur le cas du « langage du calcul des classes », à titre d'exemple paradigmatique [127].

Soit donc $\mathcal{L}$ un langage-objet contenant les éléments suivants :

(a) comme constantes, la négation « $\neg$ », la disjonction « $\vee$ », le quantificateur universel « $\forall$ » et un symbole de relation binaire « $\subset$ » exprimant l'inclusion ;

(b) un nombre infini dénombrable de variables, « x_1 », « x_2 », … ;

(c) comme symboles auxiliaires, des parenthèses « (» et «) ».

Le métalangage $\mathcal{M}$ comprend les éléments nécessaire pour décrire la structure et la syntaxe de $\mathcal{L}$ et construire des noms pour chacune de ses expressions, à savoir

1. pour chaque symbole de constante et pour les deux symboles auxiliaires de $\mathcal{L}$, un nom permettant de les désigner et que l'on notera en plaçant un trait au-dessus du symbole. Autrement dit, « $\overline{\neg}$ », « $\overline{\vee}$ », « $\overline{\forall}$ », « $\overline{\subset}$ », « $\overline{(}$ » et « $\overline{)}$ » sont des symboles de $\mathcal{M}$ nommant respectivement « $\neg$ », « $\vee$ », « $\forall$ », « $\subset$ », « (» et «) » ;
2. pour chaque variable x_i de $\mathcal{L}$, un symbole v_i qui la nomme ;
3. un opérateur de concaténation $\frown$ tel que si les lettres « φ » et « ϕ » désignent dans $\mathcal{M}$ deux expressions quelconques de $\mathcal{L}$, alors « $\varphi \frown \phi$ » désigne (dans $\mathcal{M}$) l'expression (de $\mathcal{L}$) formée par la concaténation de φ et de ψ.

Ainsi équipé, on peut nommer dans $\mathcal{M}$ toute expression du langage-objet $\mathcal{L}$. Par exemple, l'expression (de $\mathcal{L}$)

$$\text{« } \neg(x_1 \subset x_2) \text{ »}$$

est nommée (dans $\mathcal{M}$) par

$$\text{« } \overline{\neg} \frown \overline{(} \frown \ v_1 \frown \overline{\subset} \frown v_2 \frown \overline{)} \text{ »}$$

127. Originellement, tout ce qui suit est évidemment dû à Tarski. Mais nous nous sommes aussi très grandement inspirés ici de l'excellente présentation simplifiée qu'en a proposé Cozic (2009).

Nous sommes également en mesure de définir dans $\mathcal{M}$ ce que Tarski (1935) appelle l'ensemble des « fonctions propositionnelles » et qu'on nomme de nos jours l'ensemble des formules de $\mathcal{L}$:

Définition 1. *i. Pour tous i, j entiers naturels, $\overline{(} \frown v_i \frown \overline{\subset} \frown v_j \frown \overline{)}$ est une formule ;*

ii. si ϕ est une formule, alors $\neg \frown \phi$ est une formule ;

iii. si φ, ϕ sont deux formules, alors $\overline{(} \frown \varphi \frown \overline{\vee} \frown \phi \frown \overline{)}$ est une formule ;

iv. si ϕ est une formule, alors $\overline{\forall} \frown v_i \frown \phi$ en est également une.

Un énoncé se définit alors comme une formule sans variable libre. Néanmoins, nous l'avons déjà mentionné, la méthode récursive qui consiste à définir la vérité d'un énoncé à partir de la vérité des énoncés plus simples qui le composent ne peut être appliquée directement, pour la bonne et simple raison que les énoncés de $\mathcal{L}$ n'ont pas toujours pour constituants immédiats des énoncés. Par exemple, l'énoncé « $\forall x_1(x_1 \subset x_1)$ » [128] est composé à partir de la formule ouverte « $(x_1 \subset x_1)$ » par composition avec le quantificateur « $\forall$ ». La sous-formule « $(x_1 \subset x_1)$ » n'est pas elle-même un énoncé susceptible d'être vrai ou faux. La solution élaborée par Tarski va consister à introduire une autre notion sémantique, qui s'appliquera en toute généralité aux formules (formules ouvertes *et* énoncés), et à partir de laquelle on pourra définir la vérité proprement dite. Cette autre notion sémantique, c'est la notion de *satisfaction*.

La satisfaction se présente comme une relation entre les objets (ou plutôt, nous allons le voir, les suites infinies d'objets) dont parle le langage et certaines expressions du langage, plus particulièrement les formules ouvertes [129]. Voyons de quoi il retourne à partir de quelques exemples. Considérez une formule ouverte de $\mathcal{L}$ à deux variables libres, disons « $\phi(x_2, x_{11})$ ». Telle qu'elle, cette formule n'est ni vraie ni fausse. Elle ne pourra prendre une valeur de vérité que lorsqu'on aura attribué une valeur aux variables x_2 et x_{11}. Supposons qu'on attribue à ces deux variables deux objets de l'univers, disons o_2 et o_{11}, tels que lesdits objets o_2 et o_{11} —dans cet ordre— sont bien dans la relation exprimée par la formule « $\phi(x_2, x_{11})$ ». On dira alors que le couple d'objet (o_2, o_{11}) satisfait la formule. Pour prendre un exemple plus concret directement issu de notre langage-objet $\mathcal{L}$: examinons la formule « $x_2 \subset x_{11}$ ». Si on attribue l'objet o_2 à la

128. Pour plus de lisibilité, nous nous autorisons ici les guillemets ; mais bien entendu cet énoncé sera noté $\overline{\forall} \frown v_1 \frown \overline{(} \frown v_1 \frown \overline{\subset} \frown v_1 \frown \overline{)}$ dans $\mathcal{M}$.

129. Tarski insiste sur le fait que dans la mesure où la satisfaction relie des objets du monde et des expressions du langage, il s'agit bien d'une notion *sémantique*.

variable x_2 et l'objet o_{11} à la variable x_{11} et que l'objet o_2 est bien inclus dans l'objet o_{11}, alors on dira que le couple (o_2, o_{11}) satisfait la formule en question. [130]

Bien entendu, le nombre de variables libres apparaissant dans une formule ouverte, bien que fini pour chaque formule, n'est pas borné et peut donc être arbitrairement grand. Pour définir en toute généralité la relation de satisfaction entre les objets dont parle le langage et les formules de ce langage, on ne peut donc pas se limiter à la donnée d'un objet, ni même aux couples, aux triplets ou aux n-uplets d'objets. Il faudra recourir à des suites infinies d'objets. Comme $\mathcal{L}$ contient un nombre dénombrable de variables x_i ($i \in \mathbb{N}^*$), on pourra se limiter aux suites infinies dénombrables d'objets :

$$s = \langle o_1, o_2, \dots \rangle$$

Une telle suite détermine une (unique) assignation de valeurs aux variables de $\mathcal{L}$: pour tout entier i non nul, l'assignation $s = \langle o_1, o_2, \dots \rangle$ attribue l'entité o_i à la variable x_i. Pour s une telle assignation, on notera s_i l'objet attribué par s à la variable x_i. Avec les notations de $\mathcal{M}$, on obtient alors la définition récursive de la satisfaction suivante :

Définition 2 (définition récursive de la satisfaction pour $\mathcal{L}$)**.**

i. *s satisfait* $\overline{(} \frown v_i \frown \overline{\subset} \frown v_j \frown \overline{)}$ *ssi* s_i *est inclus dans* s_j *;*

ii. *s satisfait* $\overline{\neg} \frown \phi$ *ssi s ne satisfait pas* ϕ *;*

iii. *s satisfait* $\overline{(} \frown \varphi \frown \overline{\vee} \frown \phi \frown \overline{)}$ *ssi s satisfait* φ *ou s satisfait* ϕ *;*

iv. *s satisfait* $\overline{\forall} \frown v_i \frown \phi$ *ssi toute suite qui diffère de s au plus à la i-ème position satisfait* ϕ*.*

Notons que, dans cette définition, nous avons utilisé des termes comme « ne … pas », « … ou … », « toute … », « … est inclus … » qui sont censés être, dans le métalangage, les traductions de « ¬ », « ∨ », « ∀ », « ⊂ », du langage-objet [131].

130. Et pour prendre un exemple encore plus concret, on dira que la formule ouverte

$$x_2 \textit{ est plus grand que } x_{11}$$

est satisfaite par le couple d'objet (o_2 = La tour Eiffel, o_{11} = Notre Dame).

131. Attention à ne pas confondre, dans $\mathcal{M}$, les noms des éléments syntaxiques de $\mathcal{L}$ et leurs traductions. Par exemple « $\overline{\neg}$ » est un nom dans $\mathcal{M}$ désignant le symbole « ¬ » de $\mathcal{L}$, tandis « ne … pas » est une traduction dans $\mathcal{M}$ de ce qui est désigné dans $\mathcal{L}$ par « ¬ ». Autrement dit « ne … pas » est une expression de $\mathcal{M}$ ayant la même signification (à savoir la négation) que l'expression « ¬ » de $\mathcal{L}$.

Un autre exemple pour clarifier encore la situation : si mon langage-objet $\mathcal{L}$ est (une variante à structure spécifiée non sémantiquement close de) l'anglais et que mon métalangage $\mathcal{M}$ est (une variante....) du français, alors $\overline{s} \frown \overline{n} \frown \overline{o} \frown \overline{w}$ est un nom dans $\mathcal{M}$ du $\mathcal{L}$-terme « snow », tandis que le $\mathcal{M}$-terme « neige » en est une traduction.

Il est ensuite aisé de montrer que pour tout *énoncé* « ϕ » de $\mathcal{L}$, c'est-à-dire pour toute formule de $\mathcal{L}$ ne contenant aucune variable libre, deux cas de figure seulement sont possibles : ou bien toute assignation satisfait « ϕ », ou bien aucune ne le satisfait. On définit alors la vérité de la manière suivante :

Définition 3 (définition de vrai-dans-$\mathcal{L}$)**.**

$$\textit{(Def) } Vr(x) \leftrightarrow \textit{toute assignation } s \textit{ satisfait } x.$$

Il est également possible de montrer que cette définition satisfait bien la CONVENTION **T**. En outre il s'agit bien d'une définition (explicite) en bonne et due forme : le *definiendum* « $Vr(x)$ » est bien mis en équivalence avec un *definiens* dans lequel le terme défini est absent. En d'autres termes cette définition réduit la notion de vérité à celle de satisfaction.

Enfin, comme l'illustre l'exemple paradigmatique du langage de la théorie des classes, il est à peu près clair que, pour un langage dont le lexique de base est fini, la satisfaction peut être définie au moyen d'une définition récursive dans laquelle *aucun terme sémantique* n'apparaît. Voyez la définition ci-dessus pour une illustration. Cette dernière définition peut ensuite elle aussi être transformée en définition explicite dès lors que le métalangage est essentiellement plus riche que le langage-objet. Le projet tarskien semble donc mené à son terme.

1.1.4.4 Tarski et le déflationnisme

Nous avons déjà eu l'occasion de l'évoquer : les travaux de Tarski ont suscité d'innombrables commentaires et donné naissance à une littérature foisonnante. Il ne saurait être question d'en faire ici le compte-rendu exhaustif. Nous nous contenterons de rappeler les principales discussions soulevées par l'interprétation des résultats de Tarski qui nous semblent les plus pertinentes pour une histoire du déflationnisme.

Avant d'en venir à des questions d'ordre plus strictement philosophique, nous voudrions commencer par exposer certains autres résultats techniques obtenus par Tarski en marge de son élaboration d'une définition de la vérité. Ces résultats concernent les possibilités d'axiomatiser la vérité sans passer par une définition explicite. Ils ont une importance cruciale pour une discussion du déflationnisme et nous les retrouverons et

les analyserons bien plus en détails dans la suite de ce travail [132]. Néanmoins, dans la mesure où Tarski est le premier à les avoir introduits et discutés, il n'est que justice de les évoquer dès à présent.

Nous avons rappelé la rupture fondamentale qu'a représenté le *Post-Scriptum* dans l'évolution des conceptions tarskiennes de la vérité. Avant l'ajout de ce texte, Tarski aboutissait à un résultat d'indéfinissabilité absolue pour le prédicat de vérité des langages d'ordre infini [133] ; avec le *Post-Scriptum* ce résultat devient caduc et la vérité définissable pour tout langage-objet $\mathcal{L}$ fixé quel que soit son ordre fini ou infini [134], quoiqu'elle reste indéfinissable dans $\mathcal{L}$ lui-même et qu'il faille pour obtenir la définition recourir à un métalangage « essentiellement plus riche ». Nous avons également souligné que la nécessité d'user d'un métalangage d'ordre supérieur résultait en partie de l'impératif que Tarski s'était fixé d'obtenir une définition explicite de la vérité et que si l'on relâchait cette exigence d'autres possibilités s'ouvraient, notamment celle de se contenter d'une axiomatisation du prédicat de vérité. Or justement, Tarski lui-même a exploré cette possibilité. En particulier, avant la rédaction du *Post-Scriptum*, face à ce qui lui apparaissait alors comme l'impossibilité absolue de définir la vérité pour certains langages, Tarski envisage de recourir à une axiomatisation d'un prédicat « Vr » introduit comme un symbole primitif [135]. Une telle possibilité est permise en vertu du résultat suivant :

Théorème 4. Tarski *(1935, théorème III, p. 250) Si la classe de toutes les théorèmes de la métathéorie est non contradictoire et si nous adjoignons à la métathéorie le symbole « Vr » comme nouveau terme primitif et tous les énoncés décrits dans [la* convention **T***] comme de nouveaux axiomes, alors la classe des théorèmes de la métathéorie ainsi élargie sera également non contradictoire.*

Ainsi, supposons que nous partions d'un langage-objet $\mathcal{L}$ et d'une théorie objet exprimée dans ce langage [136]. Si besoin [137], on l'étend pour obtenir un métalangage et une

132. Voyez en particulier les discussions autour de l'argument de la conservativité où les approches axiomatiques évoquées ici à la suite de Tarski jouent un rôle central.
133. selon la notion d'ordre (et de richesse essentielle) pré-*Post-Scriptum* qu'il employait alors.
134. selon la nouvelle notion d'ordre (et de richesse essentielle) élaborée par Tarski dans le *Post-Scriptum*.
135. Notons que ce type d'approches était parfaitement classique pour les mathématiques de l'époque, au point qu'Hilbert parlait de « définitions implicites » des concepts par axiomatisation (en un sens relâché puisque de telles « définitions » ne remplissaient pas les hypothèses du théorème de Beth, sauf à pouvoir être transformées en définitions explicites).
136. c'est-à-dire d'un langage au sens de Tarski
137. Un tel besoin peut être superflu lorsque la théorie objet exprimée dans $\mathcal{L}$ suffit à coder sa propre

métathéorie $\mathcal{M}_0$ qui contiendront simplement les éléments nécessaires pour développer une théorie de la syntaxe de $\mathcal{L}$ (c'est-à-dire, disons, de quoi former un nom $\ulcorner\varphi\urcorner$ pour chaque énoncé φ du langage $\mathcal{L}$) et une traduction φ^* (éventuellement homographique) de chaque énoncé de $\mathcal{L}$. Supposons également que cette métathéorie « minimale » $\mathcal{M}_0$, qui pour l'instant ne contient aucune notion sémantique soit cohérente, alors la métathéorie $\mathcal{M}$ obtenue en adjoignant à $\mathcal{M}_0$ un nouveau prédicat primitif « Vr » et comme nouveaux axiomes pour en gouverner l'usage, la collection de toutes les instances de la CONVENTION **T** :

$$Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi^*,$$
$$\text{où } \varphi \text{ est un énoncé de } \mathcal{L},$$

sera elle aussi cohérente.

La démonstration de ce résultat esquissée par TARSKI (1935, p. 250) est particulièrement intéressante pour notre propos. En réalité cette démonstration établit non seulement le résultat de cohérence relative de $\mathcal{M}$ par rapport à $\mathcal{M}_0$, mais, à y regarder de près, elle établit également un résultat un peu plus fort, à savoir la conservativité de $\mathcal{M}$ sur $\mathcal{M}_0$ [138]. Ce résultat a joué un rôle central dans un débat houleux qui a opposé les partisans du déflationnisme en matière de vérité à leurs détracteurs. Nous aurons donc l'occasion d'y revenir longuement dans la suite de ce travail [139].

Toutefois, en dépit des garanties de cohérence qu'elle offre, Tarski ne paraît pas se satisfaire de cette axiomatisation de la vérité réduite aux seules **T**-équivalences . Selon lui en effet une telle axiomatisation n'est pas suffisamment forte. Il est certes trivial de vérifier qu'elle satisfait la CONVENTION **T**, par construction, mais elle ne permet pas de prouver certaines propriétés qui paraissent intuitivement inhérentes à la notion de vérité et dont on est en droit d'attendre qu'une théorie satisfaisante de la vérité puisse les démontrer. Plus précisément, Tarski affirme que la valeur du résultat de cohérence précédent est

syntaxe (nous songeons évidemment ici au langage de l'arithmétique et aux théories exprimées dans ce langage telles que l'arithmétique de Robinson $\mathcal{Q}$ ou l'arithmétique de Peano PA).

138. La démonstration de Tarski se trouve dans TARSKI (1935, p. 250-251). Voyez plus loin **??** pour une démonstration détaillée.

139. La question de la conservativité de la vérité sur une théorie-objet est centrale dans la querelle qui a opposé Ketland et Shapiro aux déflationnistes. Nous examinons cette question en détails dans la seconde partie de notre travail (chapitres ....). Signalons qu'à travers le critère d'harmonie globale, les résultats de conservativité ont également leur importance pour la partie de notre travail portant sur la question de la logicité de la vérité (*cf.* chapitre....).

> affaiblie considérablement par le fait que les axiomes indiqués dans le théorème III possèdent une force déductive très faible : la théorie de la vérité fondée sur eux ne saurait être qu'un système hautement incomplet [140], auquel manqueraient *les lois, de nature générale, les plus importantes et les plus fécondes.* (Tarski, 1935, p. 251, nous soulignons)

De fait, cette métathéorie restreinte aux seules **T**-équivalences ne peut démontrer aucun des résultats suivants [141] :

1. (Le tiers-exclu TE) exprimé sous la forme : pour tout x si x désigne un énoncé du langage-objet, alors $Vr(x)$ ou $Vr(neg(x))$ [142].
2. (Le Principe de non contradiction NC) exprimé sous la forme : pour tout x si x désigne un énoncé du langage-objet, alors $\neg Vr(x)$ ou $\neg Vr(neg(x))$.

Or ces résultats les « plus importants et féconds » sont indispensables à une théorie de la vérité satisfaisante ; c'est du moins ce que suggère Tarski.

Dans la rapide discussion de ces résultats qu'il propose, Tarski (1935, p. 251) remarque que la théorie axiomatique proposée permet de démontrer toutes les *instances* de ces « théorèmes généraux » ; c'est-à-dire que pour tout énoncé φ *fixé* du langage-objet, l'axiomatisation composée de l'ensembles des **T**-équivalences permet, par exemple, de dériver « $Vr(\ulcorner\varphi\urcorner)$ ou $Vr(\ulcorner\neg\varphi\urcorner)$ » qui est une instance de (TE). Mais elle ne permet pas de prouver l'énoncé universellement quantifié correspondant, *i.e.* le principe général (TE) lui-même [143]. Pour reprendre une expression de Tarski lui-même, le principe général représente pour ainsi dire le « produit logique infini » de ses instances, mais on ne peut le dériver à partir de ces seules instances, du moins si on s'en tient aux règles de déduction logiques habituelles. Il faut souligner ici que la source de l'insatisfaction de Tarski réside apparemment dans une contrainte d'adéquation plus forte que la seule convention **T**. Visiblement, aux yeux de Tarski, une théorie de la vérité satisfaisante

140. Le terme « incomplet » est à prendre en un sens non technique ici, puisque la « complétude » ou plutôt l' « incomplétude » évoquée l'est non pas au sens où le système d'axiomes manque de « décider » tout énoncé du langage considéré, mais simplement au sens où le système d'axiomes échoue à établir tous les résultats qu'on en attend.

141. La liste est non exhaustive. Outre les principes du tiers exclu et de non contradiction, Tarski évoque également d'autres « lois de nature générale » ; par exemple celle selon laquelle les conséquences d'énoncés vrais sont toujours des énoncés vrais (*cf.* Tarski (1935, p. 251)).

142. où $neg(x)$ est une fonction qui à tout nom d'un énoncé associe le nom de la négation de cet énoncé. Ainsi, $neg(\ulcorner\varphi\urcorner) = \ulcorner\neg\varphi\urcorner$.

143. Cette faiblesse déductive de la théorie axiomatique donnée par l'ensemble des **T**-équivalences est à la base d'une critique adressée aux déflationnistes par Gupta (*cf.* Gupta (1993)). Nous aurons l'occasion d'y revenir par la suite.

doit non seulement satisfaire la CONVENTION **T** *mais encore* montrer les « importants et fructueux théorèmes » énoncés ci-dessus : pour Tarski la CONVENTION **T** semble donc être un critère d'adéquation minimal *nécessaire* mais *non suffisant*. Là encore, nous retrouverons la question de ces critères d'adéquation supplémentaires, au delà de la seule CONVENTION **T**, dans la suite de notre travail.

Ajouter comme axiomes *ad hoc* indépendants les principes de non-contradiction (NC) et de tiers-exclu (TE) ne semble guère satisfaisant aux yeux de Tarski en raison du caractère irrémédiablement arbitraire et contingent d'un tel élargissement [144]. Pour pallier ces défauts, Tarski examine donc la possibilité d'introduire une règle de déduction supplémentaire. Il s'agit d'une forme d'ω-règle, qu'il baptise règle d' « induction infinie ». Cette règle consiste essentiellement à autoriser la déduction de la généralisation universelle d'une formule à partir de la collection infinie de ses instances [145]. Voici la façon dont Tarski lui-même l'énonce :

> si la formule donnée contient comme unique variable libre le symbole « x » appartenant à la même catégorie sémantique que les noms des expressions, et si chaque énoncé, obtenu à partir de cette formule par la substitution à la variable « x » du nom décrivant la structure d'une expression quelconque, compte parmi les théorèmes de la métathéorie, alors l'énoncé obtenu à partir de l'expressions « *pour tout x, si x est une expression, alors p* », en y

144.

> [...] tout élargissement semblable de l'ensemble considéré d'axiomes aura, semble-t-il, un caractère contingent lié à des facteurs non essentiels, tels que l'état actuel de nos connaissances dans ce domaine. En tout cas, divers critères objectifs auxquels on voudrait recourir pour le choix d'axiomes complémentaires se révèlent entièrement inapplicables. (TARSKI, 1935, p. 215)

145. Précisément, l'ω-règle habituelle en arithmétique est la règle suivante :

$$\frac{\varphi(\overline{0}), \varphi(\overline{1}), \varphi(\overline{2}), \dots}{\forall x \varphi(x)}$$

La règle suggérée par Tarski (et réduite aux énoncés) s'énoncerait comme ceci :

$$\frac{\varphi(\ulcorner \psi_0 \urcorner), \varphi(\ulcorner \psi_1 \urcorner), \varphi(\ulcorner \psi_2 \urcorner), \dots}{\forall x (En(x) \rightarrow \varphi(x))}$$

où $\psi_0, \psi_1, \dots$ parcourent l'ensemble des énoncés du langage-objet et où $En(x)$ est un prédicat désignant l'ensemble des énoncés du langage-objet.

Notons que lorsqu'on prend pour langage-objet un langage contenant suffisamment d'arithmétique et que l'on y code la syntaxe au moyen d'une arithmétisation à la Gödel, on peut faire en sorte que la règle d'induction infinie proposée par Tarski et l'ω-règle arithmétique soient équivalentes.

> substituant au symbole « p » la formule examinée, peut être jointe aussi aux théorèmes de la métathéorie. (Tarski, 1935, p. 252)

Cependant, Tarski lui-même le souligne, cette règle diffère radicalement des règles de déduction habituelles de la logique classique, notamment parce qu'elle est « essentiellement infinitaire » (ses prémisses sont constituées d'un ensemble infini de formules). Au point qu'

> on peut sérieusement douter de la compatibilité de cette règle avec la méthode déductive conçue comme on l'a fait jusqu'ici. (Tarski, 1935, p. 253)

Ceci étant, elle permet effectivement d'augmenter considérablement la force déductive des systèmes d'axiomes. Elle permet en particulier de rendre l'arithmétique de Peano complète [146]. Si on adopte cette règle, alors les seuls axiomes évoqués dans le théorème III de Tarski (1935), c'est-à-dire très exactement les instances de la convention **T**, sont suffisants pour développer une théorie de la vérité qui satisfera le critère d'adéquation renforcé avancé par Tarski : ces axiomes + l'ω-règle prouvent les plus intéressants et féconds théorèmes généraux tels que (NC) ou (TE). Malheureusement, l'adoption d'une telle ω-règle nous fait sortir du cadre de la logique standard du premier ordre. Le coût principal de cette manœuvre est que le système renforcé ainsi obtenu perd la propriété de compacité. Or, cette propriété est utilisée dans la démonstration du théorème III. Par conséquent, comme Tarski lui-même le fait remarquer, si l'on adopte cette « règle de déduction infinie », le théorème III devient indémontrable. Autrement dit, la cohérence du système élargi par les **T**-équivalences ne peut plus être prouvée. Cette solution ne semble donc pas satisfaisante aux yeux de Tarski, ce qui n'est pas étonnant pour un auteur dont le but originel [147] était, rappelons-le, de fournir une théorie formelle de la vérité qui garantisse un usage logiquement cohérent et scientifiquement légitime du prédicat de vérité.

Outre l'axiomatisation d'un prédicat « Vr » donnée par les seules **T**-équivalences , et l'axiomatisation renforcée au moyen d'un « règle d'induction infinie », il existe cependant un troisième système axiomatique pour la vérité que, curieusement, Tarski n'a pas examiné alors qu'il l'avait pourtant sous les yeux. Il s'agit de l'axiomatisation de la vérité obtenue en adjoignant à une théorie-objet munie d'une théorie de sa syntaxe, un symbole

146. Cette fois ci au sens technique du terme : tout énoncé du langage est soit démontré soit réfuté par le système ainsi renforcé. Cette règle fait également de l'arithmétique de Peano un système d'axiomes catégorique au sens où il n'admet qu'un unique modèle à isomorphisme près.
147. ou l'un des buts originels. Voyez la note 71.

de relation de satisfaction « $Sat(x, y)$ [148] » pris comme primitif et, comme axiomes pour en gouverner l'usage, l'ensemble des clauses récursives du type de celles tirées de la définition récursive donnée plus haut [149]. Ce type de théorie axiomatisée de la vérité, bien qu'elle n'ait pas été directement mentionnée par Tarski, est depuis connue sous le nom de théorie (récursive) *tarskienne* de la vérité et fait partie de la « zoologie » ordinaire des théoriciens actuels de la vérité. Les propriétés de ce genre d'extension ont donc été abondamment étudiées dans la littérature logique et philosophique contemporaine [150]. Signalons simplement ici qu'une extension de ce type permet généralement de prouver « les lois, de nature générale, les plus intéressantes et les plus fécondes » évoquées par Tarski (et non pas seulement leurs instances). Sous certaines hypothèses supplémentaires, elle permet même de prouver la cohérence de la théorie-objet [151]. En revanche, il n'est généralement pas possible de prouver la cohérence du système étendu sur la seule base de l'hypothèse de cohérence de la théorie-objet [152]. Ce type d'axiomatisation a également

148. La satisfaction est une relation binaire entre des suites d'objets (des assignations) et des formules.

149. Voyez la définition récursive de la satisfaction page 70. En reprenant *mutatis mutandis* notre exemple du langage de la théorie des classes comme langage-objet, on obtiendrait comme *axiomes* récursifs pour la satisfaction :

i. $\mathrm{Sat}(s\ ,\ \overline{(}\frown v_i \frown \overline{\subset} \frown v_j \frown \overline{)})$ ssi s_i est inclus dans s_j ;

ii. $\mathrm{Sat}(s\ ,\ \overline{\neg}\frown \phi)$ ssi non $\mathrm{Sat}(s,\ \phi)$;

iii. $\mathrm{Sat}(s\ ,\ \overline{(}\frown \varphi \frown \overline{\vee} \frown \phi \frown \overline{)})$ ssi $\mathrm{Sat}(s\ ,\ \varphi)$ ou $\mathrm{Sat}(s\ ,\ \phi)$;

iv. $\mathrm{Sat}(s\ ,\ \overline{\forall}\frown v_i \frown \phi)$ ssi pour toute suite s' qui diffère de s au plus à la i-ème position $\mathrm{Sat}(s'\ ,\phi)$.

À strictement parler, on axiomatise donc ici la satisfaction, la vérité étant ensuite définie de la manière habituelle :

$$\text{(Def) } Vr(x) \leftrightarrow \text{pout toute assignation } s,\ \mathrm{Sat}(s,\ x).$$

Néanmoins, sous certaines hypothèses (c'est le cas en particulier si on prend pour théorie-objet l'arithmétique de Peano PA), le détour par la satisfaction est inutile et on peut directement donner les clauses récursives pour un prédicat de vérité primitif « Vr ». Voyez la suite de ce travail pour un traitement technique plus détaillé de ce cas.

150. Voyez par exemple Feferman (1991), Kaye (1991) et Kotlarski (1991) et pour un panorama récent Halbach (2014).

151. Pour illustration : si PA est la théorie-objet et qu'on l'étend au moyen des clauses récursives pour la vérité *et qu'on étend* les schémas d'induction au nouveau vocabulaire, le système étendu $\mathcal{M}$ prouve la cohérence de PA :

$$\mathcal{M} \vdash Con(PA).$$

Voyez la suite de ce travail pour plus de détails.

152. En vertu du second théorème d'incomplétude de Gödel :

$$\mathcal{M} \nvdash Con(\mathcal{M})$$

et puisque

$$\mathcal{M} \vdash Con(PA),$$

joué un rôle essentiel, en tant que théorie rivale d'une théorie purement décitationnelle, dans le débat opposant Ketland et Shapiro aux déflationnistes. Nous aurons donc l'occasion de l'examiner bien plus en détails dans la suite de ce travail et d'en proposer une discussion plus approfondie ainsi qu'une comparaison avec les axiomatisations limitées aux seules **T**-équivalences [153].

Nous en venons à présent au problème spécifique de l'interprétation philosophique des travaux de Tarski et de leurs rapports avec le déflationnisme. Nous ne tenterons pas de dresser ici un panorama complet de toutes les voies selon lesquelles l'influence, immense, de Tarski s'est exercée. Nous nous concentrerons sur les liens de ses travaux avec le déflationnisme. Sur cette question les commentateurs et exégètes de Tarski sont partagés : certains veulent voir dans Tarski un précurseur du déflationnisme, voire un déflationniste avant l'heure [154], quand d'autres considèrent que Tarski était partisan d'une conception « robuste » de la vérité assimilable à une forme de théorie de la correspondance. Il faut dire que les propres (et rares) commentaires de Tarski sur le sens philosophique de son travail semblent parfois tirer dans des directions opposées.

Selon un épisode resté fameux, lors de leur rencontre à Vienne, à la lecture des épreuves du manuscrit de (Tarski, 1935), Popper fut immédiatement convaincu que ce dernier avait

> réhabilité la théorie de la correspondance de la vérité absolue ou objective. (Popper, 1963, p. 223)

De nos jours, cette interprétation « maximaliste » n'a plus guère de partisans [155]. Toutefois, on retrouve de nombreux commentateurs actuels qui considèrent Tarski comme un correspondantiste. Voici quelques références récentes défendant ce point de vue : Fernández-Moreno (2001), Niiniluoto (1999), Schantz (1998), Sher (2004, 1999) et Weingartner (1999).

À l'opposé du spectre, bien des auteurs déflationnistes contemporains se sont réclamés de Tarski : Quine (1999, § 6, p. 55) attribue à Tarski le développement classique de l'idée

---

il suit

$$\mathcal{M} \nvdash Con(PA) \rightarrow Con(\mathcal{M}).$$

153. Voyez la partie II de ce travail.
154. et avant l'introduction du terme « déflationnisme » lui-même
155. Pour une analyse critique de la lecture poperrienne des travaux de Tarski et de ses manquements, voyez Rouilhan (2007).

selon laquelle affirmer que « Brutus a assassiné César » est vrai revient à à affirmer que Brutus a assassiné César. Pour Leeds (1978, p. 120-122) les **T**-équivalences suffisent à axiomatiser la notion de vérité et c'est à Tarski que revient le mérite de l'avoir découvert. Horwich (1982) estime que ce qui distingue les déflationnistes des partisans du réalisme métaphysique est le fait qu'ils considèrent que le Schéma-**T** de Tarski

> est bien suffisant pour saisir le concept de vérité (Horwich, 1982, p. 182)

et que la signification ce dernier est « épuisée » par la caractérisation qu'en donne Tarski au moyen des **T**-équivalences (Horwich, 1982, p. 191). Enfin, Devitt (2001, p. 73) affirme que

> [b]ien que Tarski semble s'être considéré lui-même comme un théoricien de la correspondance, il est aujourd'hui généralement admis, je pense, que la théorie qu'il a réellement présenté est déflationniste.

Pour s'en tenir aux seuls écrits de Tarski en personne, à l'appui d'un Tarski correspondantiste, on trouve les nombreuses déclarations de l'auteur lui-même. En effet, Tarski affirme, à de multiples reprises, que sa conception entend saisir de manière rigoureuse l'essence de la théorie de la correspondance. Nous avons déjà cité des passages en ce sens tirés de Tarski (1944) [156]. Mais dès les premiers écrits de Tarski sur la vérité, on trouve également des déclarations similaires :

> Je me limite donc à souligner que dans toute cette étude je ne cherche qu'à saisir les intuitions exprimées par la conception dite *classique* de la vérité (selon laquelle « vrai » signifie la même chose que « correspondant à la réalité ») [...] Tarski (1935, p. 160)

tandis que quelque trente cinq ans plus tard, Tarski déclare encore

> Le caractère sémantique du terme « vrai » est clairement révélé par l'explication proposée par Aristote [N.D.T. dont Tarski se réclame] et par certaines formulations qui seront données plus tard dans cet article. On parle parfois de théorie de la vérité-correspondance en tant que théorie fondée sur la conception classique. Tarski (1969, p. 63)

Autrement dit, tout au long de sa carrière, Tarski n'a jamais varié et a toujours présenté son travail sur la vérité comme héritier d'une conception correspondantiste.

156. *Cf.* les citations données p. 40 *sq.*

Pour autant, d'autres extraits des écrits de Tarski semblent le rapprocher du courant déflationniste en matière de vérité. C'est le cas lorsqu'il déclare que les **T**-équivalences de la CONVENTION **T** sont autant de « définitions partielles » de la notion de vérité [157] et qu'

> [i]l n'y a donc pas à exiger de la définition générale d'énoncé vrai beaucoup plus que de remplir les conditions ordinaires de rectitude méthodologique et d'embrasser toutes les définitions partielles de ce type [N.D.T. *i.e.* toutes les **T**-équivalences] en tant que cas particuliers, d'en être en quelque sorte *le produit logique* (TARSKI, 1935, p. 191, nous soulignons)

De même lorsqu'il insiste sur le caractère de neutralité du concept de vérité :

> En fait, la définition sémantique de la vérité n'implique rien concernant les conditions sous lesquelles un énoncé tel que (1) :
>
> (1) *La neige est blanche*
>
> peut être affirmé. Elle implique seulement que lorsque nous admettons ou rejetons cet énoncé, nous devons être prêts à affirmer ou à rejeter l'énoncé corrélatif (2) :
>
> (2) *L'énoncé « neige est blanche » est vrai.*
>
> Ainsi pouvons-nous accepter la conception sémantique de la vérité sans abandonner nos positions épistémologiques quelles qu'elles soient. Nous pouvons demeurer réalistes naïfs, réalistes critiques ou idéalistes, empiristes ou métaphysiciens, comme nous l'étions avant. *La conception sémantique de la vérité est entièrement neutre par rapport à ces attitudes.* (TARSKI, 1944, p. 295, nous soulignons)

Cette importance centrale donnée au **T**-équivalences , le fait de les assimiler à une définition de la notion de vérité, tout comme l'insistance sur la neutralité du concept de vérité quant aux querelles métaphysiques et épistémologiques sont très proches des thèmes portés par les déflationnistes contemporains.

En revanche, il existe un point important sur lequel Tarski semble se démarquer fortement des déflationnistes. Cela concerne la place de la vérité au sein de notre entreprise de connaissance. Tarski était, semble-t-il, persuadé qu'un emploi fécond des notions sémantiques était possible. Une fois le concept de vérité clarifié et précisé au moyen d'une

157. Par exemple *in* TARSKI (1944, p. 273) ou bien *in* TARSKI (1935, p. 191).

définition, il devait être possible de l'utiliser pour obtenir de nouvelles explications de certains phénomènes ou pour prouver de nouveaux théorèmes. Nous en donnons deux exemples, dus à Tarski lui-même.

Dans le champ des sciences déductives tout d'abord, Tarski écrit :

> En ce qui concerne la possibilité d'application de la sémantique aux sciences mathématiques et à leur méthodologie, c'est-à-dire à la métamathématique, [...] nous sommes à même de signaler les résultats concrets déjà obtenus.
>
> Bien que l'on continue à douter que la notion d'énoncé vrai — en tant qu'elle est distincte de celle d'énoncé susceptible de preuve — puisse avoir quelque signification pour les disciplines mathématiques et tenir quelque place dans une discussion méthodologique de ces sciences, il me semble tout de même que cette notion de vérité constitue la plus précieuse contribution de la sémantique à la métamathématique. *Nous sommes déjà en possession d'une série de résultats métamathématiques intéressants obtenus à l'aide de la théorie de la vérité.* Ces résultats concernent les relations entre la notion de vérité et celle de démonstrabilité ; ils établissent de nouvelles propriétés de cette dernière notion (laquelle est, comme on le sait, l'une des notions métamathématiques de base) ; ils jettent aussi une certaine lumière sur les problèmes fondamentaux de la consistance et de la complétude. (Tarski, 1944, p. 303, nous soulignons)

Les résultats auxquels Tarski fait allusion ici concernent la possibilité de donner une preuve sémantique de cohérence d'une théorie-objet donnée [158], et plus généralement l'établissement d'une distinction nette entre les notions de prouvabilité formelle et de

158. preuve dont Tarski lui-même reconnaît néanmoins qu'elle est généralement d'une portée épistémique limitée :

> Ainsi la théorie de la vérité nous fournit-elle une méthode générale pour prouver la consistance dans le domaine des disciplines mathématiques formalisées. On peut cependant facilement comprendre que la preuve de consistance obtenue au moyen de cette méthode ne peut nous convaincre ou renforcer notre conviction selon laquelle la discipline considérée est effectivement consistante, que si nous réussissons à définir la vérité au moyen d'un métalangage qui ne contienne pas comme partie le langage-objet [...]. Car dans ce cas seulement les suppositions déductives du métalangage peuvent être intuitivement plus simples et plus évidentes que celles du langage-objet, quand bien même la condition de la « richesse essentielle » serait formellement remplie. (Tarski, 1944, p. 285, note 15)

vérité d'un énoncé [159].

Mais l'utilité des notions sémantiques n'est pas cantonnée aux seules mathématiques. Dans le domaine de la science empirique également, la notion de vérité peut être employée de manière féconde. Pour Tarski en effet,

> [l]a question se pose de savoir si la sémantique peut être de quelque secours pour résoudre les problèmes généraux et pour ainsi dire classiques de la méthodologie. [...]
>
> L'un des principaux problèmes de la méthodologie de la science empirique consiste à établir les conditions auxquelles une théorie ou une hypothèse empirique devrait être considérée comme acceptable. [...] [I]l me semble qu'il existe un important postulat auquel on peut très raisonnablement soumettre les théories empiriques acceptables et qui contient la notion de vérité (Tarski, 1944, p. 300-301)

Le postulat en question est donné par Tarski sous la forme suivante :

> *Dès que nous réussissons à montrer qu'une théorie empirique contient (ou implique) des énoncés faux, elle ne peut plus être considérée comme acceptable.*
>
> (Tarski, 1944, p. 302)

Tarski fait ensuite suivre ce postulat d'une discussion des raisons pour lesquelles nous sommes enclins à rejeter une théorie empirique dès lors que nous avons la preuve de son inconsistance. Pour Tarski, la vraie raison de notre attitude est que

> nous savons (ne serait-ce qu'intuitivement) qu'une théorie inconsistante contient nécessairement des énoncés faux [160]. Et nous ne sommes pas enclins à accepter une théorie dont on pourrait montrer qu'elle contient de tels énoncés [161].

159. Ce qui, dans le contexte des années trente, constituait une clarification théorique majeure.

160. On voit ici l'importance du principe (NC) de non contradiction de l'ensemble des énoncés vrais, que Tarski classait parmi « les lois, de nature générale, les plus importantes et les plus fécondes » (Tarski, 1935, p. 251) et dont il attendait qu'une théorie sémantique puisse les établir.

161. En vertu du postulat que nous venons de citer dans le corps du texte.

Anticipant sur la suite de notre travail, signalons dès à présent que dans les débats qui ont fait suite à la querelle opposant Shapiro et Ketland aux déflationnistes, certains partisans du déflationnisme ont défendu l'idée qu'il est possible de montrer qu'une théorie acceptable doit être cohérente sans faire aucune hypothèse sur sa vérité, ni même recourir à quelque notion sémantique que ce soit. Manifestement pour Tarski l'ordre des raisons est exactement inverse : c'est parce qu'une théorie dont on a montré l'inconsistance contient nécessairement des énoncés faux qu'elle devient inacceptable et doit être rejetée. Pour une comparaison voyez la suite de notre travail, en particulier les discussions qui ont suivi le débat opposant Tennant à Ketland.

Ainsi, Tarski insiste sur l'importance et la fertilité des notions sémantiques pour la méthodologie des sciences et les études fondationnelles. Cette démarche mise en avant par Tarski lui-même semble être diamétralement opposée à celle qui guide les déflationnistes, qui ne veulent voir dans la vérité qu'une notion purement expressive et sans contenu explicatif propre.

Bref, comme on peut le constater, la nature des rapports de Tarski avec le déflationnisme et la question de savoir s'il peut être considéré comme un « déflationniste avant l'heure » ou un auteur « pré-déflationniste » reste assez difficile à trancher. Plusieurs conclusions semblent possibles et les discussions restent ouvertes parmi les spécialistes de Tarski [162]. Mais quoi que l'on pense de la position de Tarski lui-même, il est toutefois incontestable que ses travaux sur la vérité ont exercé une influence primordiale sur les auteurs déflationnistes qui sont venus après lui. En lavant le concept de vérité du soupçon d'incohérence qui pesait sur lui, en montrant le rôle central de **T**-équivalences et en envisageant la possibilité d'axiomatiser ce prédicat, Tarski a ouvert la voie aux auteurs déflationnistes en matière de vérité et leur a fourni les outils techniques nécessaires sur lesquels ils ont pu s'appuyer pour développer leurs propres thèses philosophiques. Dans la section qui suit, nous allons donc nous tourner vers quelques-uns des principaux auteurs déflationnistes contemporains.

1.2 Le décitationnisme et les approches déflationnistes contemporaines

Nous allons à présent exposer une forme de déflationnisme qui ne souscrit pas au projet d'élimination du prédicat de vérité au moyen d'outils logiques tels que la quantification substitutionnelle, ni à la nécessité d'en fournir une définition explicite. Nous voulons parler ici des conceptions déflationnites contemporaines qui s'appuient sur une analyse décitationnelle du prédicat de vérité, entendue dans un sens large [163]. Le décitationnisme est sans conteste la forme de déflationnisme la plus célèbre et la plus populaire

162. Signalons pour finir sur ce point que selon Patterson (2012), les difficultés d'interprétation et les tensions apparentes opposant divers passages des écrits de Tarski ne seraient dues qu'à une mauvaise compréhension du projet global de Tarski et à une confusion entre la conception générale que Tarski se faisait de la vérité et la (ou les) définition(s) formelle(s) qu'il donne du prédicat de vérité pour tel ou tel langage formalisé fixé. *Cf.* Patterson (2012, Chapitre 5, en particulier p. 140-143).

163. C'est-à-dire qu'elles incluent le déflationnisme à la Horwich, qu'on pourrait qualifier de « décitationnisme propositionnel ».

depuis Quine. La caractéristique essentielle de ce courant de pensée est sans doute qu'il considère le prédicat de vérité avant tout comme un instrument logico-syntaxique *indispensable* (donc certainement pas éliminable ou redondant) de décitation. Mais si le statut ou la forme syntaxique du prédicat « vrai » sont conservés, l'usage de ce prédicat sera considéré comme « ontologiquement non-engageant » : la vérité n'est pas une propriété métaphysiquement substantielle qui serait attribuée à des porteurs. Le prédicat « vrai » est bel et bien indispensable et il agit bien syntaxiquement comme un prédicat, mais c'est avant tout un outil linguistique. N'étaient certaines caractéristiques ou imperfections de nos langages, sans doute serait-il superflu. Dès ses origines, avec Quine, le décitationnisme a été très influencé par les travaux de Tarski et en particulier par l'importance donnée à la CONVENTION **T** et aux **T**-équivalences . Il reste donc proche de la logique classique du premier ordre dans son énonciation et se présente souvent comme une axiomatisation formalisée de la notion de vérité. Nous présentons ici trois des principales théories déflationnistes contemporaines.

1.2.1 Quine et le décitationnisme

Quine est généralement considéré comme le père fondateur du « décitationnisme », ou théorie « décitationnelle » de la vérité [164]. À ce titre, il a exercé une influence fondamentale sur le déflationnisme aléthique contemporain. Les réflexions de Quine sur la notion de vérité prennent place parmi les thèmes centraux de sa pensée : la thèse du holisme de la confirmation empruntée à Duhem et radicalisée sous la forme du holisme sémantique ; le naturalisme et le refus de toute philosophie première surplombant ou fondant le discours scientifique ; le rôle central de la logique et du langage dans le compte de nos engagements ontologiques ; et bien sûr, l'indétermination de la traduction et l'inscrutabilité de la référence qui fournissent à Quine des arguments pour rejeter hors du discours scientifique légitime les notions de signification, de synonymie et d'analyticité. Nous ne reviendrons pas ici en détails sur ces thèses célèbres et parfois controversées, en tout cas toujours abondamment discutées [165], nous nous concentrerons pour l'essentiel

164. C'est en particulier à lui qu'on doit le slogan :

« la vérité est décitation »

Cette formule, frappante et demeurée célèbre, se trouve textuellement dans nombres de ses écrits, par exemple dans (QUINE, 1987, p. 265) ou dans (QUINE, 1990, p. 117).

165. Nous nous permettons de renvoyer le lecteur aux nombreuses études consacrées à la philosophie de cet auteur majeur du siècle passé, notamment : en anglais, GOCHET (1986) et HYLTON (2007), en

sur ce que Quine a dit à propos de la vérité.

Tout au long de l'évolution de sa pensée, Quine est resté fidèle à sa conception décitationnelle de la vérité. Si cette dernière est développée de la manière la plus détaillée et la plus approfondie dans Quine (1970, en particulier au chapitre 1), on la retrouve dans une formulation quasi inchangée vingt ans plus tard (Quine, 1990, chapitre 5), tandis que les schémas décitationnels pour la vérité, la satisfaction et la dénomination étaient déjà mis en avant dès Quine (1953, p. 190).

Comme on pouvait s'y attendre de la part de cet amoureux des paysages désertiques, la théorie quinienne de la vérité obéit à un double souci d'économie ontologique et de parcimonie méthodologique. Quine l'introduit parfois comme une reprise critique et une mise à distance d'une forme naïve de vérité correspondance, ou comme « le résidu valide » [166] de la conception de la vérité comme correspondance une fois qu'on l'a débarrassée de la phraséologie philosophiquement suspecte qui l'accompagne ordinairement. Le passage suivant tiré de Quine (1990) illustre bien ce mode de présentation :

> Les véhicules de la vérité étant déterminés, en quoi consiste alors leur vérité ? Ils se qualifient comme vrais, nous dit-on, en correspondant à la réalité. La correspondance mot à mot, néanmoins, ne peut faire l'affaire ; elle invite à encombrer paresseusement la réalité d'une troupe bizarre d'objets fictifs, aux seules fins de la correspondance. Une idée plus nette consiste à poser des *faits*, correspondant aux énoncés vrais pris comme des touts ; ce n'est à nouveau qu'une mystification. Des objets en grand nombre, concrets et abstraits, sont à coup sûr nécessaires pour rendre compte du monde ; mais les faits n'apportent aucune contribution au-delà de leur appui spécieux à une théorie de la correspondance.
>
> Pourtant comme Tarski nous l'a appris, il y a un fond de vérité dans la théorie de la vérité-correspondance. Au lieu de dire :
>
> « la neige est blanche » est vrai si et seulement si c'est un fait que la neige est blanche,
>
> nous pouvons effacer « c'est un fait que » comme étant vide, et du même coup les faits eux-mêmes :
>
> « la neige est blanche » est vrai si et seulement si la neige est blanche.

français, Gochet (1978), Hookway (1992), Laugier (1992), Monnoyer (2006) et Rivenc (2008).

166. Quine (1990, p. 132)

> Attribuer la vérité à l'énoncé, c'est attribuer la blancheur à la neige ; telle est la correspondance dans cet exemple. L'attribution de la vérité se borne à effacer les guillemets. La vérité est la décitation. (Quine, 1990, p. 116-117, italiques de l'auteur) [167]

On perçoit bien ici le glissement opéré par Quine : il congédie la notion de fait et la relation de correspondance, laquelle se trouve réduite —ou remplacée ?— par ce qui ressemble fort à une **T**-équivalence .

Cette équivalence, tout comme l'exemple de la blancheur de la neige sont évidemment repris de Tarski [168]. Toutefois, Quine en donne une lecture assez personnelle. Et pour cerner l'austère frugalité de la théorie décitationnelle, il nous faut encore prêter attention aux éléments composant l'équivalence mise en avant par Quine. Il nous faut en particulier jeter un œil sur ce que doivent être pour Quine les porteurs de vérité, ces entités auxquelles le prédicat « vrai » en censé s'appliquer. À ce sujet, ce n'est pas un hasard si les textes que Quine consacre à la vérité s'ouvrent par une attaque en règle de la notion de proposition entendue comme signification des énoncés [169]. Quine y reprend ses arguments célèbres [170] : les théories affrontent en bloc le tribunal de l'expérience et il faut abandonner toute idée de siginification individuelle ou de contenu propositionnel associé à un énoncé pris isolément. Contrairement à Frege ou Ramsey par exemple, Quine rejette donc toute notion de référence propositionnelle attribuée aux énoncés. Mais si la vérité ne s'applique pas aux propositions, quels peuvent être les « véhicules » de la vérité ? Comme indiqué dans l'extrait ci-dessus Quine considère que le prédicat de vérité

167. Voyez aussi Quine (1987, p. 265) pour un exposé très similaire :

> [...] On nous dit que la vérité de « la neige est blanche » est due au fait que la neige est blanche et que l'énoncé vrai « la neige est blanche » correspond au fait que la neige est blanche. L'énoncé « la neige est blanche » est donc vrai si et seulement si c'est un fait que la neige est blanche. Ainsi, tout factice et illusoire qu'il soit, le fait se trouve acculé dans un coin où il ne nous reste plus qu'à lui donner le coup de grâce. L'expression « c'est un fait que » est vide et peut être rejetée ; « c'est un fait que la neige est blanche » se réduit à « la neige est blanche ». Dès lors, notre compte rendu de la vérité de « la neige est blanche » en termes de faits se ramène donc à ceci : « la neige est blanche » est vrai si et seulement si la neige est blanche.

168. Cette filitation est d'ailleurs fréquemment revendiquée par Quine. Outre le passage cité ci-dessus, voyez aussi, par exemple Quine (1970, p. 22) et Quine (1987, p. 265).

169. Voyez Quine (1970, chapitre 1, p. 9 à 21) et Quine (1990, chapitre V, § 32).

170. Dont la genèse remonte à sa critique de l'analyticité adressée aux empiristes logiques (Quine, 1954, 1935, 1951) et qui furent sans doute portés à leur plus haut degré de développement et de radicalité dans Quine (1960).

doit s'appliquer directement aux énoncés eux-mêmes [171], dont les conditions d'individuation lui semblent moins problématiques. Ce choix des énoncés comme porteurs de vérité semble le rapprocher de Tarski [172]. Mais là encore, par rapport à la construction tarskienne dont il se revendique, Quine effectue un subtil déplacement. Dans la fameuse CONVENTION **T** telle qu'elle a été énoncée par Tarski, la définition d'un prédicat « Vr » pour être matériellement adéquate en tant qu'analyse du concept de vérité pour un langage $\mathcal{L}$ devait impliquer toutes les **T**-équivalences tarskiennes, c'est-à-dire outes les instances du schéma-T suivant :

(T) X est Vr ssi p

où « X » est remplacé par un nom [173] ou une *description structurelle* d'un énoncé de $\mathcal{L}$ alors que « p » est remplacé par une *traduction* de cet énoncé dans le métalangage $\mathcal{L}'$ de la métathéorie employée pour parvenir à une définition de la vérité (pour $\mathcal{L}$). Chez Quine, le schéma ci-dessus prend la forme de ce qu'on pourrait appeler un schéma de décitation :

(SD) « p » est vrai ssi p

où, bien sûr, « p » est remplacé par le *même* énoncé à gauche et à droite du si et seulement si ; à gauche entre guillemets (autrement dit en position de citation) et à droite en position d'usage. Évidemment, pour parler de *dé*citation il est impératif que les noms d'énoncés apparaissant dans l'équivalence soient obtenus par citation, c'est-à-dire, disons, par une mise entre guillemets. Ce détail n'en est peut-être pas un et le glissement vers une lecture décitationnelle des **T**-équivalences n'est pas anodin. Contrairement à ce proposait Tarski, il n'y a plus chez Quine de traduction dans un métalangage expressivement plus riche, et les descriptions structurelles [174] des énoncés deviennent des énoncés cités entre guillemets. Il est alors possible de présenter la prédicat de vérité comme un simple outil syntaxique permettant d'effacer les guillemets. Et c'est exactement ce que fait Quine :

171. Soyons grammaticalement précis : les énoncés seront les porteurs ou les véhicules de la vérité mais c'est bien sûr aux noms qui les désignent que le prédicat « vrai » lui-même sera apposé.
172. Dans TARSKI (1935), Tarski attribue lui aussi la vérité aux énoncés. Et c'est une définition d'un prédicat de vérité pour les énoncés d'un langage formel qu'il entend fournir.
173. de quelque forme que ce soit.
174. Les noms par description structurelle (*structural descriptive names*) introduits et employés par Tarski dans TARSKI (1935).

> Le prédicat de vérité est un dispositif pour neutraliser les guillemets (QUINE, 1970, p. 25)

Par ce mouvement, Quine « trivialise » en partie les **T**-équivalences : la construction tarskienne d'une métathéorie censée permettre de rendre compte de certaines propriétés sémantiques du langage-objet et dont la CONVENTION **T** constituait le noyau est remplacée par des équivalences décitationnelles, construites autour d'une subtile distinction entre usage et mention d'un *même* énoncé, où le prédicat de vérité devient « transparent ».[175]

Ceci étant, si Quine entend attribuer la vérité directement aux énoncés dont la nature lui semble moins impénétrable que celle des propositions, il reste que l'attribution de vérité ne peut avoir de sens que si elle s'effectue sur un énoncé interprété, c'est-à-dire sur un énoncé pris à l'intérieur d'un langage (interprété) donné. En effet, une même suite de signes (phonèmes ou graphèmes par exemple) pourrait très bien apparaître comme vraie pour un locuteur d'un certain langage et fausse pour le locuteur d'un autre langage. Quine est bien conscient du problème puisqu'il déclare :

> On conçoit que par une extraordinaire coïncidence la même suite de sons

175. Pour dire quelques mots supplémentaires sur ce qui distingue la conception quinienne de celle de Tarski, examinez la **T**-équivalence suivante :

L'énoncé formé du substantif *snow* suivi du verbe *is* et de l'adjectif *white* est Vrai-dans-$\mathcal{L}$ ssi la neige est blanche

où le langage-objet $\mathcal{L}$ serait l'anglais, et où notre métalangage serait le français+quelques outils permettant de désigner les expressions de $\mathcal{L}$. Cette équivalence obéit parfaitement aux canons tarskiens. Difficile nous semble-t-il néanmoins de parler ici d'une simple décitation à propos de Vrai-dans-$\mathcal{L}$.

Mais, pourrait-on dire, les équivalences décitationnelles ne sont qu'un cas particulier, peut-être le plus fondamental, des équivalences tarskiennes dans lesquelles la traduction dans le métalangage de l'énoncé du langage objet dont il est question est la traduction dite homophonique (où p est traduit par lui-même). Certes, mais deux points demeurent :

1. D'un point de vue (méthodo)-logique, la présence ou non d'une traduction, même homophonique ou homographique, est cruciale. Ce n'est pas la même chose par exemple de manipuler syntaxiquement une suite de graphèmes pour leur accoler des ornements comme des guillemets ou des parenthèses que l'on peut ensuite effacer à loisir, et de traduire un énoncé. Cet aspect du problème est d'autant plus important pour quelqu'un comme Quine, puisqu'on connaît le soupçon qu'il a jeté sur la notion de traduction interlinguistique – nous renvoyons à ses fameuses analyses sur la situation de traduction radicale qui se trouve sous-déterminée par les données empiriques et sur l'inscrutabilité de la référence dans QUINE (1960).
2. Cela laisse entière la question de la manière dont sont obtenus les noms d'énoncés au moyen de guillemets de citation. Le statut de ces noms nous semble loin d'être aussi trivial que Quine le laisse entendre. Mais nous aurons l'occasion de revenir sur ce point.

Pour une comparaison plus approfondie du projet de Tarski et de sa réinterprétation quinienne, voyez RIVENC (1996).

> ou de caractères pourrait servir pour « $2 < 5$ » dans une langue et pour « $2 > 5$ » dans une autre. Lorsque nous parlons de « $2 < 5$ » comme d'un énoncé éternel[176] , nous devons entendre que *nous le considérons exclusivement comme un énoncé de notre langue* et que nous n'affirmons que sont vrais que ceux de ses signes concrets qui sont des énonciations verbales ou des inscriptions produites à l'intérieur de notre communauté linguistique. (Quine, 1970, p. 26-27, nous soulignons)

Ceci conduit Quine à considérer que la vérité est, selon ses propres termes, une notion purement *immanente* : le prédicat de vérité ne peut, *in fine*, être appliqué par un locuteur qu'à un énoncé de son *propre* langage, ou à la limite celui de sa communauté linguistique[177]. Et comme pour Quine (apprendre à) parler un langage c'est déjà adopter une théorie du monde, l'attribution de vérité n'aura au fond de sens qu'à l'intérieur d'un cadre théorique et linguistique donné :

> C'est plutôt lorsque nous nous replaçons au cœur d'une théorie qui existe réellement, et qui est acceptée au moins à titre d'hypothèse, que nous pouvons parler, et parler avec sens, de tel ou tel énoncé comme d'un énoncé vrai. S'il y a un sens à appliquer le qualificatif de « vrai » à un énoncé, c'est à un énoncé exprimé dans les termes d'une théorie donnée, et considéré du point de vue de cette théorie, complète avec les réalités que cette théorie « pose ». (Quine, 1960, § 6, p. 55, traduction modifiée)[178]

Cette immanence de la vérité cadre parfaitement avec le naturalisme quinien. Pas plus qu'il n'y a de réalité appréhendable hors et indépendamment de notre schème conceptuel, pas plus n'y a-t-il de point de vue transcendant permettant de juger, pour ainsi dire de l'extérieur, quelle est notre meilleure théorie du monde. Et les attributions de vérité doivent donc toujours se faire à l'intérieur du cadre théorique du locuteur qui réalise l'attribution. C'est pourquoi, Quine finira par conclure qu'

176. Quine appelle énoncés éternels les énoncés dont les éléments indexicaux, déictiques, termes vagues et autres ont été trivialisés de façon à ce que l'énoncé puisse se voir attribuer une valeur de vérité « éternelle ». Par exemple, « Il pleut » n'est pas un énoncé éternel, tandis que « il pleuvait à Paris le 12 juillet 1967 à 12 h 32 » s'en rapprocherait plus. Voyez Quine (1970, chapitre 1, § Signes concrets et énoncés éternels, p 25-28)

177. Les attributions de vérité à des énoncés d'une langue étrangère ne sont de ce point de vue que secondaires et ne se font que modulo une traduction de ces énoncés dans mon propre idiome, traduction qui est elle-même relative à un manuel de traduction donné.

178. Par souci de cohérence et d'uniformité avec les autres extraits traduits de l'anglais cités ici, nous avons rendu le terme anglais original « *sentence* » par « énoncé » plutôt que par « phrase » comme le proposent Dopp et Gochet dans leur traduction.

> [a]ppeler un énoncé vrai, ce n'est que l'inclure dans notre propre théorie du monde. (Quine, 1995, § IV Vérité : immanente ou transcendante ?, p. 353) [179]

Mais si Quine ne veut voir dans la vérité qu'une notion immanente ne servant qu'à effacer l'usage des guillemets de citation accolés aux énoncés de notre propre langage, quelle utilité peut bien avoir le prédicat « vrai » ? Ne faudrait-il pas plutôt le laisser de côté comme on l'a fait avec les propositions ? La réponse apportée par Quine à cette question constitue précisément l'héritage décisif de la théorie décitationnelle pour le déflationnisme contemporain.

À l'instar de Ramsey et des partisans de la théorie de la vérité-redondance, Quine adhère à la théorie de la disparition du prédicat « vrai » dans les cas de citations directes d'un énoncé (ou d'un ensemble fini d'énoncés) pris isolément :

179. Mettons néanmoins en garde le lecteur sur ce que Quine semble dire ici. Tout d'abord quelques lignes après le passage cité on trouve les réflexions suivantes :

> *Appeler* un énoncé vrai, ai-je dit, c'est l'inclure dans notre science, mais cela ne veut pas dire que la science fixe la vérité. Elle peut avoir tort. Nous avançons en testant notre théorie scientifique par la prédiction et l'expérience, et en la modifiant lorsque c'est nécessaire, en *quête* de la vérité. La vérité se profile ainsi tel un havre vers lequel nous ne cessons de nous diriger et de mettre le cap. Elle est un idéal de la raison pure, selon les mots de Kant. Très bien : immanente par tous ses autres aspects, transcendante par celui-ci. (Quine, 1995, § IV Vérité : immanente ou transcendante ?, p. 353)

D'autre part, d'après Quine (1960, p. 55), tout cela ne signifie pas qu'on puisse

> définir [la vérité] de manière dérivée en disant qu'un énoncé déterminé S est vrai si il (ou si une de ses traductions) appartient à θ [une systématisation totale de la science] (Quine, 1960, p. 55, trad. modifiée)

La raison donnée par Quine à cette impossibilité est, une fois encore, une conséquence du holisme sémantique :

> parce que, en général, cela n'a aucun sens d'égaler un énoncé d'une théorie θ à un énoncé S donné à part de la théorie. [...] un énoncé S est dépourvu de sens sauf relativement à sa propre théorie ; il est dépourvu de sens inter-théoriquement. (Quine, 1960, p. 55, trad. modifiée)

Plus généralement, cette nature strictement interne/immanente de la notion de vérité selon Quine n'est pas sans poser problème. Au point que certains commentateurs de Quine ont pu y voir un risque de relativisme et une menace pour le réalisme revendiqué de Quine —en particulier lorsqu'on lui ajoute la sous détermination des théories par les données empiriques. Lorsque deux théories du monde sont également (empiriquement) possibles et acceptables quoique logiquement incompatibles, que dire d'un énoncé inclus dans l'une et rejeté par l'autre ? Est-il vrai pour l'une et fausse pour l'autre ? De même, comment rendre compte de l'évolution temporelle de nos théories scientifiques ? Si un énoncé autrefois partie prenante de notre meilleure théorie du monde est aujourd'hui rejeté par la science actuelle, on ne dira pas qu'il était vrai hier bien qu'il soit faux aujourd'hui. C'est donc qu'il y a peut-être plus dans l'attribution de vérité à un énoncé que la simple incorporation à une théorie adoptée à un moment donné. Sur ce point voyez Gochet (1986, en particulier p. 107-124), ainsi que Hylton (2007, p. 276-277). Quine évoque lui-même cette difficulté et y répond laconiquement dans Quine (1960, p. 56). Pour les ultimes (mais à nos yeux là encore très elliptiques) réflexions de Quine sur cette question, voyez sa réponse à Davidson dans Quine (1995, § IV).

> Parler de la vérité d'un énoncé donné n'est qu'un détour ; nous devrions simplement énoncer cet énoncé : du même coup nous ne parlons plus du langage, mais du monde. Tant que nous n'avons à parler de la vérité d'énoncés donnés isolément, la théorie la plus parfaite de la vérité est celle que Wilfrid Sellars a appelée la théorie de la vérité évanescente [180]. (Quine, 2008, p. 22)

Dans de tels cas, le prédicat de vérité est un ornement dont on peut se dispenser. Il est redondant et éliminable. Conformément à l'analyse selon laquelle le prédicat de vérité ne fait qu'annuler l'effet des guillemets, l'énoncé, par exemple,

« La neige est blanche » est vrai

peut être remplacé par l'énoncé

La neige est blanche

sans aucune perte de contenu selon Quine.

En revanche, là où Quine se départ de Ramsey et des partisans de la théorie de la vérité-redondance, c'est dans l'analyse qu'il propose de l'utilité d'un prédicat décitationnel pour l'expression de certaines généralités. Il y a selon Quine en effet « des occasions dans lesquelles, bien qu'ayant en vue une réalité d'ordre non linguistique, nous sommes incités à procéder indirectement et à parler d'énoncés » [181]. Les occasions de ce type sont celles où nous cherchons à exprimer une généralité mais « où nous la cherchons le long de certains plans obliques, que nous ne pouvons pas embrasser en généralisant sur des objets » [182]. Quine explique plus précisément ce qu'il a en tête à travers l'analyse d'exemples :

> Nous pouvons généraliser sur « Tom est mortel », « Richard est mortel », *etc.* , sans parler de vérité ni d'énoncés : nous pouvons dire « Tous les hommes sont mortels ». Nous pouvons de même généraliser sur « Tom est Tom », « Richard est Richard », « 0 est 0 », et ainsi de suite, en disant « Toute chose est elle-même ». Quand par ailleurs nous voulons généraliser sur « Tom est mortel ou Tom est non mortel », « La neige est blanche ou la neige est non blanche », et ainsi de suite, nous nous élevons jusqu'à parler de vérité et

180. « *disappearance theory of truth* » dans le texte anglais original. L'évanescence dont il est question ici ne renvoie donc pas à une nature vague, insaisissable ou imprécise, mais désigne bien la qualité de ce qui tend à disparaître.
181. Quine, 2008, p. 23.
182. Quine, 2008, p 23.

> d'énoncés en disant « Tout énoncé de la forme 'p ou non p' est vrai », ou « Toute disjonction d'un énoncé avec sa négation est vraie ». Ce qui nous contraint à cette montée sémantique, ce n'est pas que « Tom est mortel ou Tom est non mortel » porterait de quelque façon sur des énoncés tandis que « Tom est mortel » et « Tom est Tom » porteraient sur Tom. Ces trois énoncés portent tous sur Tom. Nous ne nous élevons qu'à cause de la manière oblique dont les instances sur lesquelles nous généralisons sont reliées les unes aux autres. (Quine, 2008, p 23)

Quelle est cette manière « oblique » qui nous contraint au détour par la mention d'énoncés ? C'est tout simplement que la généralisation cherchée, contrairement à celle de « tous les hommes sont mortels », ne peut pas s'exprimer directement à l'aide d'une quantification universelle portant sur une variable objectuelle :

> Nous étions en mesure d'exprimer notre généralisation « Toute chose est elle-même » sans monter car les changements qu'on a fait sonner en passant d'une instance à une autre —« Tom est Tom », « Richard est Richard », « 0 et 0 »—étaient des changements de nom. De même, avec « Tous les hommes sont mortels ». Cette généralisation peut s'écrire « x est mortel pour tous les *hommes* x »— *i.e.* tous les objets x d'une espèce qui est telle que « Tom » est un nom de l'un d'entre eux. Or que serait l'interprétation analogue de la généralisation de « Tom est mortel ou Tom est non mortel » ? Elle s'écrirait « p ou non p pour tous les objets p d'une espèce qui est telle que des énoncés en sont des noms ». Mais les énoncés ne sont pas des noms, et cette interprétation est incohérente, car elle emploie « p » à la fois dans des places qui appellent des membres de phrase et dans une place qui appelle un substantif. Donc, pour parvenir à l'assertion générale que nous cherchons, nous montons d'une marche et nous parlons sur des énoncés : « Tout *enoncé* de la *forme* 'p ou non p' est vrai ». (Quine, 2008, p. 22-23)

L'emploi du prédicat de vérité dans les exemples ci-dessus obéit donc à une nécessité pour ainsi dire « grammaticale » : il s'agit de rendre possible l'expression de certaines généralités au moyen de la quantification objectuelle en rétablissant la valeur assertive des énoncés cités. Selon Quine, le prédicat de vérité est précisément indispensable dans ces cas où, en raison de certaines complications techniques, on ne peut utiliser directement les énoncés pour parler du monde. On peut en donner d'autres exemples qui permettront

d'illustrer un peu plus ce dont il retourne. Ainsi, quand on cherche à asserter l'ensemble des énoncés d'un certain type ou d'une certaine forme : sans le prédicat de vérité

« Toute disjonction d'un énoncé avec sa négation est vraie »

ne pourrait pas s'exprimer sans recourir à des variables (et des quantifications) propositionnelles, ou bien à des conjonctions infinies, ou peut-être à d'autres outils logiques qui nous font sortir du cadre de la logique du premier ordre habituelle et de sa quantification strictement objectuelle. De même, sans prédicat de vérité, ce qu'exprime l'énoncé

« Tous les théorèmes de l'arithmétique PA sont vrais »

demanderait pour être formulé l'emploi de conjonctions infinies (ou d'un autre outil logique équivalent). Un autre cas typique est celui de l'énonciation dite indirecte, c'est-à-dire les cas où l'on veut asserter un énoncé sans que celui-ci ne soit directement disponible. Dire :

« La dernière phrase écrite par Platon est vraie »

est une manière d'asserter le dernier énoncé écrit par Platon alors qu'on peut très bien ne pas être en position d'identifier ici et maintenant quel était au juste cet énoncé.

Quine baptise cette technique de généralisation employant la vérité et des énoncés mentionnés entre guillemets, la « montée sémantique ». Le prédicat de vérité sert en fait à nous « ramener vers le monde », à redescendre au niveau de la réalité quand un détour par un parler en termes d'énoncés se révèle indispensable. Le prédicat « vrai » élimine l'ascension sémantique en effaçant l'effet des guillemets. C'est là sa seule véritable utilité et c'est précisément la raison qui le rend indispensable. Pour Quine donc, le prédicat de vérité (n') est (qu') un outil de décitation.

Par rapport aux théoriciens de la vérité-redondance, Quine opère de ce fait un renversement de perspective : le cas des énoncés généraux qui, chez Ramsey par exemple, apparaissait comme un obstacle à l'élimination de la vérité devant être surmonté au moyen de l'introduction d'une quantification non-standard, devient au contraire la raison d'être du prédicat de vérité. C'est dans de tels cas que le prédicat « vrai » trouve son usage comme outil permettant l'expression de certaines généralités sans toucher à la nature des variables et des quantifications. [183]

183. Contrairement à Ramsey, Quine était donc hostile à toute tentative d'élimination du prédicat de vérité dans les énoncés généraux, au moyen d'outils tels que la quantification substitutionnelle. D'une part,

Ainsi, selon Quine, le prédicat de vérité, s'il a bien la forme syntaxique d'un prédicat, ne sert pas à attribuer une véritable propriété à des énoncés. Ce n'est en réalité qu'un instrument expressif, un outil servant à parler indirectement du monde, utile et indispensable dans certaines situations de généralisation. De ce point de vue, on ne doit donc pas s'attendre à ce que l'expression « vrai » désigne une notion importante nécessitant une analyse approfondie. De même, en dehors de sa fonction d'auxiliaire expressif, la vérité n'aura sans doute pas de rôle important à jouer dans nos explications du monde. Bien au contraire, si la nature de la vérité se borne à cet usage décitationnel, alors ses caractéristiques essentielles seront entièrement saisies par les équivalences décitationnelles. Hors de la mise au jour de cet usage décitationnel, il n'est donc nul besoin de réduction, de définition ou d'analyse supplémentaire du prédicat de vérité. Dans cette mesure, bien que Quine n'ait pas proposé lui-même une version entièrement formalisée de sa théorie, on peut sans doute considérer que l'ensemble des équivalences décitationnelles formaient à ces yeux une sorte d'axiomatisation satisfaisante du prédicat de vérité. C'est en tout cas ce que nous semble suggèrer le passage suivant :

> Le compte rendu décitationnel de la vérité ne définit pas le prédicat vérité – pas au sens strict de « définition » ; la définition au sens strict nous dit en effet comment éliminer l'expression définie de tout contexte souhaité en faveur d'une notation précédemment établie. Mais en un sens moins rigoureux le compte rendu décitationnel définit la vérité. Il nous dit ce que c'est qu'être vrai pour un énoncé quelconque, et il nous le dit en termes aussi clairs pour nous que l'énoncé en question lui-même. Nous comprenons ce que c'est qu'être vrai pour l'énoncé « la neige est blanche » aussi clairement que nous comprenons ce que c'est pour la neige qu'être blanche. (Quine, 1990, p. 119)

Pour résumer l'apport de Quine, nous pouvons dire qu'en insistant sur le rôle expressif central du prédicat de vérité, en l'identifiant à un simple outil logico-syntaxique de décitation, et en présentant les **T**-équivalences (dans une lecture décitationnelle) comme

parce que Quine était généralement très opposé à ce type de logique non-standard prolongeant la logique classique du premier ordre, la quantification devant toujours se comprendre comme une quantification sur des objets et permettre en dernière analyse de dresser le compte de nos engagements ontologiques. Pour lui, la quantification substitutionnelle et les variables propositionnelles, s'il est toujours possible de les introduire d'un point de vue technique, ne sont qu'un artifice et doivent précisément être analysées et définies à partir de la quantification objectuelle. D'autre part, parce que le rôle joué par le prédicat de vérité dans l'expression des généralités est précisément ce qui le rend indispensable et explique (et justifie) sa présence dans notre idiome.

une analyse (quasi) exhaustive de cette notion, Quine a ouvert la voie au déflationnisme contemporain. Tout en conservant les intuitions d'un Frege ou d'un Ramsey sur l'absence de contenu propre apporté par le prédicat « vrai » lorsqu'on l'accole à un porteur de vérité, Quine a expliqué en quoi un tel prédicat était néanmoins indispensable et inéliminable de notre langage en vertu du gain d'expressivité qu'il lui apporte. Ce rôle purement expressif suggère en outre que la vérité n'est pas une propriété importante dont la structure devrait être mise au jour par une analyse philosophique et qui pourrait jouer un rôle central [184] au sein de nos explications scientifiques.

1.2.2 La théorie minimale d'Horwich

Dans son ouvrage *Truth* paru pour la première fois en 1990, Paul Horwich a proposé ce qui constitue sans doute l'un des exposés les plus complets et systématiquement développé d'une conception déflationniste de la vérité. Son livre a suscité de nombreux commentaires et une seconde édition, dans laquelle Horwich tente de répondre à certaines critiques, est parue en 1998 (Horwich, 1998b). Cette seconde édition est accompagnée d'un *Postscript* [185] dans lequel Horwich synthétise et précise sa position. Comme nous allons le voir, ses thèses sont grandement inspirées à la fois par Quine et par les théoriciens de la redondance, tout en s'en démarquant sur des points importants.

Pour désigner ses propres conceptions déflationnistes, Horwich emploie les termes de « minimalisme en matière de vérité » et de « théorie minimale de la vérité » :

> Je nommerai ma théorie de la vérité « *la théorie minimale* », et j'appellerai les remarques qui l'accompagnent au sujet de son adéquation « *la conception minimaliste* ». (Horwich, 1998b, p. 6, italiques de l'auteur)

Dans son *Postscript*, Horwich caractérise la position déflationninste dont il se réclame de la manière suivante :

> L'attitude déflationniste envers la vérité —et sa variante particulière que j'appelle minimalisme— sont des réactions à l'idée naturelle et répandue selon laquelle la propriété de vérité possède une sorte de nature sous-jacente et que notre problème en tant que philosophes est de dire ce que cette nature est, d'analyser la vérité soit conceptuellement, soit substantiellement, et de

184. En dehors bien sûr de son rôle d'auxiliaire expressif permettant la formulation de certaines généralités

185. Le texte de ce *Postscript* est également paru, à quelques modifications mineures près, sous le titre *The Minimalist Conception of Truth* dans Blackburn et Simmons (1999).

> spécifier, au moins à gros traits, les conditions nécessaires et suffisantes pour qu'une chose soit vraie. [...] Mais [le déflationniste] conteste qu'il existe un quelconque espoir d'obtenir une définition explicite ou une analyse réductive de la vérité, même très approximatives. (Horwich, 1998b, p. 120-121, *Postscript*)

Cette conception déflationniste est ainsi dessinée par contraste avec les conceptions traditionnelles —que ce soient les théories correspondantistes, cohérentistes, vérificationnistes ou pragmmatistes de la vérité— qui attribuent à la vérité une nature profonde et mystérieuse qu'il appartiendrait aux philosophes de révéler et d'analyser. Ce scepticisme quant à l'existence d'une nature propre de la vérité n'est pas sans évoquer les thèses des théoriciens de la « redondance » de la vérité, ou théorie éliminative de la vérité. Horwich revendique d'ailleurs cet héritage puisqu'il poursuit :

> Bien entendu, ceci rappelle grandement l'ancienne « théorie de la redondance » de Frege (1891), Ramsey (1927, 1991), Ayer (1935) et Strawson (1964) : l'idée que
>
> La proposition *que p* est vraie
>
> ne signifie ni plus ni moins que simplement
>
> *p*
>
> Et la théorie de la redondance est en effet *une forme anticipée de déflationnisme*. (Horwich, 1998b, p. 122, *Postscript*, nous soulignons)

Mais s'il reconnaît que les idées déflationnistes ont connu de nombreux précurseurs [186], Horwich considère néanmoins que jusqu'à présent elles n'ont pas reçu de formulation réellement satisfaisante. En développant sa théorie minimale de la vérité il entend donc pallier les insuffisances des précédentes tentatives et répondre aux critiques qui leur ont été adressées.

186. Dans une note de bas de page, Horwich propose la généalogie suivante :

> Des vues plus ou moins déflationnistes à propos de la vérité ont été adoptées et défendues (sous des formes variées et à des degrés divers) par Frege (1891, (1918-1923)), Ayer (1935) et Ramsey (1927), Wittgenstein (1953, 1922), Strawson (1950) et Quine (1970). Ces dernières années, l'idée a été développée par Grover, Camp et Belnap (1975), Leeds (1978), Horwich (1982), Fine (1984), Field (1994a, 1986), Williams (1986), Loar (1987), et Brandom (1994, 1988).
> (Horwich, 1998b, p. 5, note de bas de page)

Tout comme Quine, Horwich choisit comme point de départ de sa réflexion sur la vérité le critère d'adéquation minimal cher à Tarski dont il reprend l'exemple canonique. Horwich affirme ainsi :

> [...] quelle que soit la théorie de la vérité que nous adoptions professionnellement, nous sommes tous prêts à inférer
>
> La croyance que *la neige est blanche* est vraie
>
> à partir de
>
> La neige est blanche
>
> et vice versa. Et, plus généralement, nous acceptons toutes les instances des « schémas de vérité » :
>
> La croyance (conjecture, assertion, supposition ....) *que* p est vraie ssi p.
>
> (Horwich, 1998b, p. 121)

Horwich met donc en avant le « schéma de vérité » suivant :

(E) La proposition que p est vraie si et seulement si p [187]

Plus précisément, il déclare :

> Les axiomes de la théorie sont les propositions telles que
>
> (1) ⟨⟨ La neige est blanche ⟩ est vraie ssi la neige est blanche ⟩
>
> et
>
> (2) ⟨⟨ Mentir est mal ⟩ est vraie ssi mentir est mal ⟩
>
> c'est-à-dire, toutes les propositions dont la structure est
>
> (E*) ⟨⟨ p ⟩ est vraie ssi p ⟩
>
> (Horwich, 1998b, p. 17)

Précisons qu'ici les symboles « ⟨ » et « ⟩ » sont introduits par Horwich comme des raccourcis syntaxiques permettant d'alléger la notation de la locution « la proposition

187. où p est une (méta-)variable censée être remplacée par une proposition.

que p » [188]. Avec cette notation, Horwich obtient le schéma de vérité suivant :

(E) $\langle p \rangle$ est vraie si et seulement si p

dont la collection infinie des instances donne les axiomes de la théorie minimale [189].

Toutefois, le point crucial de la position déflationniste développée par Horwich n'est pas tant d'accepter toutes les instances du schéma de vérité, mais réside plutôt dans l'affirmation qu'à elle seule cette collection d'axiomes suffit à remplir tous les objectifs qu'on est en droit d'exiger d'une théorie de la vérité. Dans son *Postscript* Horwich poursuit ainsi son propos :

> Mais au lieu d'adopter la conception traditionnelle selon laquelle une *analyse* de la vérité doit encore être donnée —une explication réductive plus profonde que les schémas de vérité, qui expliquera pourquoi nous en acceptons

188. Voyez la note de la page 10 :

> J'écrirai « $\langle p \rangle$ » pour « la proposition que p », et « ssi » pour « si et seulement si ». (Horwich, 1998b, p. 10, note de bas de page 4)

Curieusement, quelques pages plus loin, dans une autre note qui suit le passage que nous venons de citer dans le corps du texte, Horwich donne une explication différente :

> J'emploie la convention selon laquelle entourer n'importe quelle expression, e, avec des crochets d'angle « ⟨ » et « ⟩ », produit une expression désignant le *constituant propositionnel exprimé par* e. (Horwich, 1998b, p. 18, note de bas de page 3)

Cette variante par rapport à la note de la page 10 est sans doute introduite par Horwich pour pouvoir traiter les cas tels que l'axiome (1) ⟨⟨ La neige est blanche ⟩ est vraie ssi la neige est blanche ⟩, où, selon ses propres dires, l'axiome est construit à partir de l'énoncé « la neige est blanche » et où, semble-t-il, les crochets s'appliquent à des énoncés (même si ces derniers expriment une proposition).

La situation n'est pas très claire : les crochets « ⟨. . . ⟩ » sont-ils une simple notation commode, ou forment-ils une sorte d'opérateur ? S'ils sont des opérateurs, agissent-ils sur les propositions ou sur les expressions ? À vrai dire ce point nous est resté quelque peu obscur. Mais ce qu'il nous semble important de bien retenir est que, contrairement au décitationnisme quinien, les schémas de vérité qui constituent la théorie minimale d'Horwich sont constitués à partir des *propositions* et non pas à partir des énoncés.

189. Quelques remarques au passage s'imposent ici. Tout d'abord la collection infinie de ces axiomes ne forme pas un ensemble. Cela peut se montrer au moyen d'un argument diagonal à la Cantor. Voyez Horwich (1998b, p. 20 note de bas de page 3). De plus, si l'on ne prend pas plus de précautions ces axiomes vont donner une théorie incohérente puisqu'ils laisseront la porte ouverte au paradoxe du menteur. Il faut donc exclure certaines instances du schéma de vérité pour conserver une axiomatisation cohérente. Horwich est bien conscient du problème et il le traite à la section 10 du chapitre 2 de son ouvrage. Sans préciser la façon dont cela peut être accompli, Horwich donne les directives suivantes :

> Étant donné nos objectifs, il nous suffit d'admettre que certaines instances du schéma d'équivalence ne doivent pas être incluses comme axiomes de la théorie minimale, et de remarquer que les principes gouvernant notre sélection des instances à exclure sont, par ordre de priorité : (a) que la théorie minimale ne produise pas de contradictions du « genre de celle du menteur » ; (b) que l'ensemble des instances exclues soit aussi petit que possible ; et —peut-être tout aussi important que (b)— (c) qu'il existe une spécification constructive des instances exclues qui soit aussi simple que possible. (Horwich, 1998b, p. 42)

> les instances— le déflationniste soutient que, dans la mesure où notre engagement envers ces schémas rend compte de tout ce que nous faisons avec le prédicat de vérité, nous pouvons supposer qu'ils le définissant implicitement. Notre simple acceptation de leurs instances constitue notre compréhension de la notion de vérité. Il n'est nul besoin d'une analyse conceptuelle —ni d'une définition de la forme
>
> « vrai » signifie « F »
>
> où « F » est une expression composée de termes plus fondamentaux que le prédicat de vérité. De plus, il n'y aura pas non plus d'analyse non définitionnelle de la vérité, aussi rudimentaire soit-elle —pas de découverte substantielle de la forme
>
> La vérité de x *consiste dans* le fait que x possède la propriété F.
>
> D'où le terme « déflationnisme ». (Horwich, 1998b, p. 121-122, italiques de l'auteur)

Par son recours aux propositions, par ses équivalences tracées entre une proposistion —p— et la proposition qui lui attribue la vérité —la proposition que p est vrai—, par son refus d'accorder à la vérité une quelconque nature profonde en attente d'être analysée, la position d'Horwich n'est pas sans rappeler les réflexions d'un Ramsey (1927) ou de ses héritiers éliminativistes [190]. Néanmoins, Horwich insiste sur ce point, le minimalisme en matière de vérité n'est pas un éliminativisme. Bien au contraire, à l'instar de Quine, Horwich adresse aux théoriciens de la redondance toute une série de critiques : les premières portent sur la rôle indispensable de la vérité en tant qu'outil expressif, les secondes sur l'emploi de quantifications non-standard, les troisièmes sur le statut d'authentique propriété de la notion de vérité.

En premier lieu, pour Horwich comme pour Quine, la principale erreur des théorie de la vérité-redondance est d'avoir manqué de saisir la fonction expressive du prédicat de vérité, qui en fait toute l'utilité et le rend indispensable [191]. Ce type d'argument illustrant le rôle expressif de la vérité, de la même veine que ceux énoncés par Quine, est développé à plusieurs reprises par Horwich au cours de son ouvrage [192]. En voici une version directement adressée aux partisans de la redondance :

190. Voir les sections 1.1.2 et 1.1.3 de ce chapitre.
191. Et s'il est indispensable il n'est par conséquent évidemment ni redondant ni éliminable.
192. Voyez Horwich (1998b, chapitre 1, p. 2), Horwich (1998b, chapitre 2, section 6, p.31-33), Horwich (1998b, *Postscript*, p. 122-123).

> [...] les théoriciens de la redondance n'avaient pas grand chose à dire concernant la *fonction* de notre concept de vérité. Mais s'il est réellement redondant, pourquoi diable possédons-nous une telle notion ? Une vertu du minimalisme est de contenir une réponse satisfaisante à cette question —la même que celle qui fut d'abord proposée par Quine (1970)— à savoir, que le prédicat de vérité joue un rôle vital pour nous permettre de saisir certaines généralisations. (Horwich, 1998b, p. 122)

S'en suit une illustration de ce rôle, tout à fait semblable à celle que nous avons déjà rencontrée avec Quine :

> [...] il existe une importante classe de généralisations qui ne peuvent pas être construites de cette façon [N.D.T : Horwich fait ici référence à la quantification objectuelle universelle] : par exemple, celle dont les instances comprennent
>
> Si Florence sourit alors Florence sourit
>
> Comment pouvons-nous extraire la loi logique qu'elle instancie ? [...] La solution fournie par notre concept de vérité est de transformer chacune de ces propositions en une proposition qui lui est manifestement équivalente —mais qui peut être généralisée de manière normale. Ainsi, étant donné l'équivalence de
>
> p
>
> et de
>
> L'affirmation que p est vraie
>
> [...] nous obtenons [...]
>
> Toute affirmation de la forme « si p alors p » est vraie [...]
>
> À partir de [cela], nous pouvons déduire (moyennant les schémas de vérité) toutes les affirmations que nous souhaitions initialement généraliser. [...] Il est ainsi effectivement utile d'avoir un terme qui soit gouverné par les schémas de vérité —en dépit de leur trivialité. Il existe clairement une *raison d'être* [N.D.T : en français dans le texte] pour un concept ayant précisément les caractéristiques que le minimaliste attribue à la vérité. (Horwich, 1998b, p. 123)

À la suite de Quine, Horwich s'oppose donc aux partisans de la théorie de la redondance pour souligner le caractère indispensable du prédicat de vérité en tant qu'outil expressif de généralisation. Horwich déclare par exemple :

> En fait le prédicat de vérité existe uniquement en raison d'un certain besoin logique [...]. C'est uniquement dans ce rôle, et non pas comme le nom de quelque ingrédient stupéfiant de la nature, que le concept de vérité apparaît de façon si répandue dans la réflexion philosophique. (Horwich, 1998b, p. 4)

Ou bien encore :

> Je ne suggère pas, bien entendu, que le prédicat de vérité a été *délibérément* introduit pour remplir cette fonction utile. Mais je *fais* l'hypothèse que son utilité, telle que je viens de la décrire, est ce qui explique sa présence. Car s'il n'avait pas du tout de valeur, l'usage s'en serait vraisemblablement perdu ; et pour ce qui est des autre fonctions qu'il pourrait avoir, il n'existe tout simplement pas de candidats plausibles. (Horwich, 1998b, p. 33)

Pour Horwich, c'est donc bien sa fonction d'outil expressif permettant de formuler des généralisations qui explique la présence inéliminable du prédicat de vérité au sein de nos constructions théoriques [193].

Un second point sur lequel Horwich se range aux côtés de Quine pour critiquer les théoriciens éliminativistes de la vérité concerne le rejet des outils logiques non standards permettant d'éliminer le prédicat de vérité [194] voire d'en donner une définition [195]. À diverses reprises en effet, Horwich examine et rejette la possibilité de recourir à la quantification substitutionnelle pour formuler des généralisations sans s'appuyer sur les prédicat de vérité [196]. Selon lui, la quantification substitutionnelle

> avec ses règles syntaxiques et sémantiques spéciales serait un ajout encombrant à notre langage. L'intérêt de notre notion de vérité est qu'elle fournit une alternative simple à cet appareillage. (Horwich, 1998b, p. 32)

193. Si l'on met de côté les partisans des théories prophrastiques, cette insistance sur le rôle expressif du prédicat de vérité est d'ailleurs une constante de quasiment *tous* les auteurs déflationnistes contemporains.
194. *Cf.* les théories prophrastiques Grover, Camp et Belnap (1975), 1.1.3.
195. *Cf.* Ramsey (1991), 1.1.2.
196. Les critiques d'Horwich sur la quantification substitutionnelle se trouvent aux endroits suivants : Horwich (1998b, chapitre 1, p. 4 note de bas de page), Horwich (1998b, chapitre 2, section 6, p. 31-33), et Horwich (1998b, *Postscript*, p. 124-125).

Le prédicat de vérité a cette vertu qu'il

> nous permet d'éviter les complexités et les obscurités de la quantification substitutionnelle. (Horwich, 1998b, p. 125)

Et qu'il

> nous autorise à obtenir les effets de la généralisation substitutionnelle sur les énoncés et les prédicats, mais au moyen de variables ordinaires (*i.e.* de pronoms), qui portent sur des *objets*. (Horwich, 1998b, p. 4, note de bas de page)

On retrouve donc chez Horwich le même arbitrage que chez Quine en faveur de la quantification objectuelle standard. Cette préséance accordée à la quantification objectuelle et cette importance dévolue au prédicat de vérité dans l'expression de certaines généralisations conduisent Horwich à formuler une troisième critique envers les théoriciens de la redondance. Comme nous l'avons vu précédemment [197], certains partisans de la théorie éliminative de la vérité allaient en effet jusqu'à affirmer que la vérité n'était pas une authentique propriété [198], et proposaient même de remettre en cause la forme syntaxique prédicative de cette notion [199]. Mais si, associée à la quantification objectuelle, la vérité est censée permettre d'exprimer des généralités, il est impératif qu'elle prenne la forme d'un prédicat [200] Et il est donc impératif, du moins d'après les apparences de la syntaxe, qu'elle désigne une propriété. Horwich est donc très prudent avec l'idée selon laquelle la vérité ne serait pas une « authentique » propriété. Il s'en explique au chapitre 2 Horwich (1998b, chapitre 2, section 9, p. 37-40) et dans son *Postscript* Horwich (1998b, *Postscript*, p. 125 et section 8 p 141-143). Selon Horwich,

> le minimalisme n'induit, en lui-même, aucune réponse spécifique à la question [de savoir si la vérité est une propriété]. (Horwich, 1998b, p. 141)

Cela va dépendre de la conception que l'on se fait de ce que doit être une (authentique) propriété. Pour autant, Horwich rappelle qu'il est « *vital*[201] » pour le minimalisme que la vérité soit formalisée logiquement par un prédicat (de façon à permettre l'interaction

197. *Cf.* la section 1.1.3.
198. Et pour cause, puisqu'elle était censée pouvoir être éliminée sans perte de notre discours scientifique légitime.
199. Nous pensons tout particulièrement ici aux théories prophrastiques qui veulent traiter la vérité sous la forme de pro-phrases, sur le modèle des pro-noms, et lui dénient donc le statut de prédicat.
200. Ne serait-ce que pour pouvoir s'appliquer aux *objets* (en l'occurrence les énoncés ou les propositions) sur lesquels portent les quantifications.
201. C'est le terme employé par Horwich lui-même, Horwich (1998b, p. 141).

avec la quantification objectuelle). Selon une conception « libérale » de la notion de propriété, d'après laquelle tout terme qui fonctionne logiquement comme un prédicat renvoie à une (authentique) propriété, le minimalisme implique donc que la vérité est bien une propriété. Mais, souligne Horwich, rien n'empêche d'élaborer des conceptions plus strictes de ce qui constitue une authentique propriété dont certaines excluront la vérité [202]. Horwich donne quand même quelques indices supplémentaires sur la nature particulière de la propriété de vérité :

> « est vrai » est un prédicat français parfaitement correct — et [...] on pourrait très bien prendre cela comme critère conclusif pour représenter une propriété de *quelque* type. Ce que le minimaliste souhaite néanmoins souligner est que la vérité n'est pas une propriété *complexe* ou *naturelle* [*naturalistic*] mais une propriété d'un autre genre (Field (1992) suggère le terme de « propriété *logique* ») [...] Selon le minimalisme, nous devrions [...] nous garder d'assimiler *être vrai* à des propriétés telles qu'*être turquoise*, *être un arbre* ou *être fait d'aluminium*. Faute de quoi, nous nous trouverons en train de chercher sa structure constitutive, son comportement causal et ses manifestations typiques —caractéristiques propres à ce que j'appelle les « propriétés *complexes* » ou « *naturelles* ». Nous serons déconcertés lorsque ces attentes seront inévitablement frustrées et enclins à conclure que la vérité est profondément obscure. (Horwich, 1998b, p. 37-38, italiques de l'auteur)

Le point d'équilibre recherché par Horwich semble donc être que la vérité soit suffisamment une propriété pour être traitée syntaxiquement comme un prédicat —de manière à permettre à la vérité de jouer son rôle d'outil pour exprimer des généralisations—, mais sans être pour autant considérée comme une propriété dont la structure devrait donner lieu à une analyse réductive au delà de la donnée des schémas de vérité.

En résumé, Horwich définit donc sa théorie minimale comme étant la donnée de toutes les instances (non paradoxales) du schéma de vérité suivant

$$\text{(E) } \langle p \rangle \text{ est vrai ssi } p,$$

Horwich insiste également sur le rôle fondamental du prédicat de vérité dans l'expression de certains énoncés généraux et souligne que c'est ce rôle qui rend indispensable la notion de vérité et en constitue la véritable « raison d'être ». Simultanément, il rejette

202. Le cas limite étant ici le nominalisme intégral selon lequel il n'existe aucune propriété et seuls existent les objets.

la quantification substitutionnelle et maintient que la vérité doit être traitée comme une propriété, au moins sur le plan syntaxique. Sur tous ces points, Horwich semble en plein accord avec le décitationnisme de Quine.

Si on examine de plus près la formulation propre de la théorie minimale, ce parallèle avec le décitationnisme quinien peut même, nous semble-t-il, être poussé un cran plus loin. Dans la formulation des axiomes de sa théorie minimale, Horwich utilise les parenthèses angulaires qui sont, disons, des raccourcis syntaxiques pour la locution « la proposition que ». Les axiomes de la théorie minimale d'Horwich sont donc l'ensemble [203] des instances de

$$\text{(E) } \langle p \rangle \text{ est vraie ssi } p$$

ce qui est syntaxiquement très proche de l'ensemble des équivalences décitationnelles qui instancient le schéma

$$\text{(SD) } \text{« } e \text{ » est vrai ssi } e$$

où e est censé être remplacé par un énoncé. L'une et l'autre de ces versions sont évidemment directement héritières des **T**-équivalences tarskiennes.

Au delà de la proximité dans la forme que prennent les axiomes des deux théories (minimale et décitationnelle), le passage suivant tiré de Horwich (1998b) nous semble révélateur :

> Ce qui permet à la notion de vérité de jouer [son] rôle [dans l'expression des généralisations] est simplement que, pour tout énoncé déclaratif
>
> $$p$$
>
> nous est fourni une énoncé qui lui est équivalent
>
> $$\text{La proposition que } p \text{ est vraie}$$
>
> où l'énoncé d'origine a été transformé en un groupe nominal, « la proposition que p », occupant une position ouverte aux variables objectuelles, et où le prédicat de vérité sert uniquement à restaurer la structure d'un énoncé : il agit simplement comme un *dé-nommeur* [*de-nominalizer*] [...] pour que le prédicat de vérité remplisse sa fonction [...] *il n'est pas nécessaire de supposer quoi que ce soit d'autre à propos de la vérité*. Le rôle conceptuel et théorique

203. En un sens relâché puisqu'il s'agit en fait d'une collection et non pas d'un ensemble.

> tout entier de la vérité peut être expliqué sur cette base. (Horwich, 1998b, p.)

Nous voulons voir dans cet extrait comme un écho de la devise quinienne :

« la vérité est décitation »,

et de son affirmation :

> Le prédicat de vérité est un dispositif pour neutraliser les guillemets. (Quine, 1970, p. 25)

Si nous pouvions parler en son nom, nous serions tentés de dire que, pour Horwich,

« *la vérité est dénominalisation* »

et que son (unique) rôle est de nous ramener à la proposition d'origine quand nous avons a dû faire un détour par un nom de cette proposition. De même, nous pourrions affirmer que, pour Horwich, le prédicat de vérité est un « dispositif pour neutraliser la nominalisation ». Le parallèle avec Quine est si frappant que, par abus de langage, nous serions tentés d'appeler la théorie minimale d'Horwich un « décitationnisme propositionnel », à ceci près que la façon dont on obtient le nom d'une proposition [204] ne se fait pas par simple citation et qu'elle est, du moins en apparence, plus complexe et plus mystérieuse que l'opération qui consiste à mettre un énoncé entre guillemets pour en obtenir un nom.

Malgré ces nombreuses convergences entre le minimalisme d'Horwich et le décitationnisme de Quine, il existe un point sur lequel ce deux auteurs divergent voire s'opposent : il s'agit précisément de la question de la nature des porteurs de vérité. Dans le cadre de ce travail, nous avons délibérément laissé de côté ce problème, mais nous allons néanmoins dire quelques mots ici sur ce qu'en déclare Horwich pour bien saisir toutes les particularités de son déflationnisme.

Parmi plusieurs candidats potentiels pour endosser le rôle de porteurs de vérité, Horwich (1998b, p. 16) choisit les propositions [205]. La première raison qu'il invoque pour justifier ce choix est la conformité à l'usage commun :

204. Ce qu'Horwich note $\langle p \rangle$ et qui correspond à la locution « la proposition que p ».
205. Horwich (1998b, p. 16) donne la liste suivantes de candidats possibles :
(a) les énonciations [*utterances*]
(b) les énoncés [*sentences*]
(c)les déclarations, croyances, suppositions [*statements, beliefs, suppositions*]
(d) les propositions [*propositions*]

> Je suivrai le langage ordinaire en supposant que la vérité est une propriété des propositions (HORWICH, 1998b, p. 16)

Dans le chapitre qu'il consacre à l'examen et à la défense des propositions, Horwich développe un peu plus cette idée. Sans en donner de définition générale, HORWICH (1998b, chapitre 6) introduit la notion de propositions comme étant les objets sur lesquels portent nos attitudes propositionnelles :

> [...] chaque fois que quelqu'un possède une croyance, un désir, un espoir ou quelqu'une de ce qu'on appelle des attitudes propositionnelles, alors son état mental consiste en l'existence d'une certaine relation entre lui et un genre particulier d'entité : à savoir, la *chose* qui est crue, désirée, espérée, *etc.* (HORWICH, 1998b, p. 86)

Pour Horwich, l'intérêt fondamental de postuler l'existence de propositions est précisément de permettre une formalisation correcte de nos attributions d'attitudes propositionnelles, conforme à (une version « enrégimentée » de) l'usage ordinaire.

> Le mérite considérable de cette théorie [postulant des propositions] est qu'elle apparaît fournir une explication adéquate des propriétés logiques des attributions de croyances et autres du même genre [206]. (HORWICH, 1998b, p. 86)

Bien entendu, Horwich est conscient que cet engagement envers l'existence des propositions est controversé et il s'emploie à montrer que les arguments classiques avancés contre les propositions ne sont pas concluants [207]. Mais il ajoute également que cet engagement est

206. HORWICH (1998b, p. 86-87) s'appuie sur des raisonnements du type suivant :
(1) Oscar croit qu'il va pleuvoir
(2) Barnabé dit qu'il va pleuvoir
Donc,
(3) Oscar croit ce que Barnabé dit
Et,
(4) Il existe une chose qu'Oscar croit et que Barnabé a dite.
Ce type de raisonnement peut être formalisé au moyen des règles logiques habituelles *à condition* qu'on accepte que les termes singuliers tels que
ce qu'Oscar croit
ce que Barnabé a dit, ou encore
qu'il va pleuvoir
réfèrent bien à quelque chose, à un type d'*entités*. Libre à nous d'appeler ensuite ces entités des *propositions*.

207. Nous ne développerons pas ce point ici. Pour plus de détails sur la défense horwichienne des propositions nous revoyons au chapitre de *Truth* consacré à cette question HORWICH (1998b, chapitre

> bien moins essentiel qu'il pourrait sembler à première vue. Car il ne présuppose que fort peu concernant *la nature* des propositions. (Horwich, 1998b, p. 16)

Le minimalisme horwichien, s'il s'appuie sur la notion de propositions, laisse en effet ouverte toute possibilité quant au statut des ces entités [208]. Une seconde raison de ne pas trop d'inquiéter de ce parti-pris méthodologique est qu'il est en grande partie optionnel : Horwich (1998b, p. 98-103) montre comment, moyennant certains principes auxiliaires, on peut dériver la notion de vérité pour les énonciations [*utterances*] à partir de celle de vérité pour les propositions, et *vice-versa*. On peut donc choisir l'une ou l'autre comme point de départ, la conception minimaliste ne s'en trouve pas chamboulée :

> Le langage ordinaire suggère que la vérité est une propriété des propositions et que les énonciations, croyances, assertions, *etc.* , hérite leur caractère de vérités de leurs relations aux propositions. Cependant, les dérivations ci-dessus montrent que cette façon de voir les choses n'a pas de mérite explicatif particulier. La conception de la vérité pour chaque type d'entités est tout autant minimaliste. Et en supposant n'importe laquelle d'entre elles, on peut facilement en dériver les autres. (Horwich, 1998b, p. 102, voir également le *Postscript*, p. 133-135)

S'il préfère les propositions qui lui semblent plus commodes, Horwich suggère donc que le minimalisme peut tout aussi bien s'appuyer sur une notion de vérité pour les énonciations [*utterances*] ou pour les énoncés [*sentences*].

Malgré cela, la conception minimaliste de la vérité renferme tout de même une conséquence fondamentale pour la théorie des propositions —et plus généralement pour la nature de la signification. Horwich le souligne très clairement :

> Le minimalisme implique que la notion de proposition ne dépende pas de la notion de vérité. Car, pour le minimaliste, la direction de priorité conceptuelle va dans le sens inverse : dans la mesure où notre concept de vérité est constitué par notre acceptation des instances « La propositions *que p* est vraie ssi *p* », nous devons déjà préalablement être en capacité de comprendre

6 : *Propositions et énonciations*) ainsi que aux autres ouvrages dans lesquels Horwich développe et approfondit sa position : Horwich (1998a, 2005, 2010).

208. Dans le chapitre 6 de son ouvrage, Horwich semble toutefois favorable à une versions double, « œcuménique » de la notion de proposition qui allie les propositions comme objets russeliennes et les propositions frégéennes construites à partir des significations abstraites.

> les propositions. (Horwich, 1998b, p. 16)

De même, dans le *Postscript* :

> [...] le minimalisme est principalement la thèse selon laquelle les schéma d'équivalence est conceptuellement fondamental vis-à-vis du prédicat de vérité [...]. Ceci implique que les divers concepts vériconditionnels [209] soient postérieurs au concept de proposition, et par conséquent qu'il soit possible de posséder le concept de proposition sans posséder le concept de vérité. (Horwich, 1998b, p. 130)

Ainsi, le déflationnisme horwichien exclut toute possiblité d'une théorie vériconditionnelle de la signification en général, et des propositions en particulier. C'est une conséquence qu'Horwich accepte fort volontiers et qui l'amène, entre autres raisons, à privilégier une théorie non-vériconditionnelle de la signification comme usage :

> Ces conséquences de la perspective minimaliste sont justifiées par la théorie de la signification comme usage [210]. (Horwich, 1998b, p. 131)

Pour conclure sur Horwich, la thèse centrale de sa conception minimale de la vérité est que l'axiomatisation donnée pas les instances du schéma

$$\text{(E)}\ \langle p \rangle \text{ est vraie ssi } p$$

suffit à donner une analyse exhaustive du concept de vérité et à rendre compte de tous les usages légitimes de cette notion. Cette axiomatisation conserve la nature prédicative de la notion de vérité lui permettant de remplir le rôle expressif qui la rend indispensable pour la formulation de certaines généralisations. S'appuyant sur la notion de proposition pour éclaircir la notion de vérité, cette théorie minimale a également pour conséquence que la notion de proposition —ou plus généralement celle de signification d'un porteur de vérité— doit être comprise préalablement et indépendamment de la notion de vérité. Sur ce dernier point Horwich s'accorde avec un autre partisan important du déflationnisme contemporain : Hartry Field.

1.2.3 Le déflationnisme méthodologique de Field

Aux côtés d'Horwich, Hartry Field est sans nul doute l'autre représentant majeur du déflationnisme actuel. Bien que ces deux auteurs s'accordent pour prôner une attitude

209. En premier lieu la vérité elle-même.
210. Horwich développe sa propre version d'une théorie de la signification comme usage dans Horwich (1998a, 2005, 2010).

déflationniste envers la vérité —ou plus généralement envers les notions sémantiques traditionnelles comme la référence, la satisfaction ou la dénotation— ils ont néanmoins des divergences quant à la manière dont cette attitude doit être développée [211]. Le déflationnisme de Field est d'autant plus remarquable que ce philosophe a dans un premier temps été un ardent défenseur d'une forme de théorie de la vérité-correspondance [212]. Dans un livre paru en 2001, *Truth and the Absence of Fact*, Field reprend les articles de sa période correspondantiste, en leur adjoignant des postscripts, ainsi que des articles plus tardifs dans lesquels il développe sa position déflationniste. Cet ouvrage offre donc une vue générale des changements de doctrines de Field en matière de vérité. Si celles-ci ont connu une évolution radicale, une sorte de retournement à 180°, il est néanmoins frappant de constater que Field est demeuré fidèle tout au long de son parcours à certains principes méthodologiques fondamentaux. Les travaux de Field s'inscrivent en effet, tout comme ceux de Quine, dans la postérité critique de l'empirisme logique et souscrivent ouvertement au naturalisme méthodologique et plus particulièrement au physicalisme. Pour le dire rapidement, le naturalisme méthodologique postule une continuité forte, tant dans les buts que dans les méthodes, entre les sciences et la philosophie. Les outils et les concepts de la philosophie doivent donc être *a minima* compatibles avec les canons d'une bonne méthodologie scientifique [213]. L'une des conséquence ou des variantes possibles de ce principe méthodologique est la doctrine physicaliste [214]. Le physicalisme, pour le dire là encore très rapidement, postule que les seuls objets et les seules propriétés existant réellement dans le monde sont les objets et les propriétés physiques [215]. Un

211. Pour une comparaison rapide mais utile par l'un des deux protagonistes, on pourra consulter Field (1992).

212. Voyez en particulier Field (1972) ainsi que Field (1978, 1974, 1973) pour un exposé des thèses correspondantistes du jeune Field.

213. Tout ceci est évidemment assez vague. Pour plus de précisions sur le naturalisme, qu'il soit méthodologique ou ontologique, et sur l'importance de ce paradigme devenu ultra-dominant dans l'univers philosophique anglo-saxon depuis la seconde moitié du vingtième siècle, voir Papineau (2016).

214. Le physicalisme et le naturalisme vont très souvent de pair, mais pas systématiquement. Par exemple, les arguments d'indispensabilité dits de Putnam-Quine s'appuient sur une forme de naturalisme méthodologique pour défendre l'existence d'objets mathématiques, c'est-à-dire d'objets non physiques, ce qui s'oppose à une (version) stricte du physicalisme. Sur les rapports entre physicalisme et naturalisme, voir à nouveau Papineau (2016).

215. Le terme même de « physicalisme » a été introduit par Neurath dans les années 1930 et était revendiqué, comme une thèse anti-métaphysique, par les membres du Cercle de Vienne. On peut aussi voir le physicalisme comme un avatar moderne du matérialisme, au point que les deux termes sont souvent employés de manière interchangeable. Sur les diverses versions et les diverses formulations du physicalisme, voire Stoljar (2017).

autre de ses thèmes centraux est la notion de clôture causale[216] du monde physique : les seuls objets ou propriété entrant dans des lois causales permettant d'expliquer des phénomènes physiques ne peuvent eux-mêmes n'être que des objets ou des propriétés physiques[217]. Par conséquent, tous les phénomènes, même ceux qui ne semblent pas être proprement physiques à première vue[218] doivent pouvoir être expliqués *in fine* en termes de faits, de lois et d'objets uniquement physiques[219]. Pour pouvoir comprendre l'évolution intellectuelle qui a conduit Field à abandonner ses thèses correspondantistes pour embrasser le déflationnisme au point d'en devenir l'un des représentants les plus éminents, il faut bien garder à l'esprit ce cadre méthodologique que nous venons de rappeler et auquel Field est resté attaché sans discontinuité. Avant d'en venir aux positions proprement déflationnistes du Field actuel, nous commençons par revenir brièvement sur ses anciennes conceptions correspondantistes et sur les buts théoriques qu'elles étaient censées remplir à ses yeux. Cela s'avère nécessaire pour bien saisir le point de départ des réflexions de Field sur la vérité.

Dans les années 1970, lorsque Field publie ses premiers travaux sur la vérité, une part importante des discussions agitant les cercles de philosophie analytique portait sur la question de savoir si les notions sémantiques et les diverses démarches théoriques qui semblaient s'appuyer sur elles étaient compatibles avec les canons acceptables d'une bonne méthodologie scientifique. Field (1972) propose une contribution à ce débat sous la forme d'une critique et d'un développement de la théorie tarskienne de la vérité. Selon Field, contrairement à ce que pouvaient laisser croire certaines déclarations de Tarski lui-même, la définition tarskienne du concept de vérité pour un langage-objet fixé ne permet pas d'établir la compatibilité de ce concept avec le physicalisme. En effet, les clauses récursives de la définition tarskienne montrent bien comment la valeur de vérité d'un énoncé complexe dépend des valeurs de vérité de ses constituants. Ce faisant, elles permettent effectivement de réduire la vérité à ce que Field baptise la « dénotation primitive », c'est-à-dire à la référence et à l'extension des termes primitifs du langage

216. Sur l'importance de cet argument pour le physicalisme, voire encore Papineau (2016) et Stoljar (2017).

217. À titre d'exemple d'une doctrine incompatible avec le physicalisme, citons le dualisme cartésien qui postulait l'existence d'une substance mentale distincte de la matière et néanmoins capable de produire des effets physiques.

218. On peut penser ainsi aux phénomènes mentaux, si tant est que de tels phénomènes existent.

219. Une variante stricte du physicalisme demandera que chaque propriété soit réduite à une propriété physique ; une variante plus faible exigera simplement la « survenance » de toute propriété sur les propriétés strictement physiques. Voir à nouveau Stoljar (2017) ainsi que McLaughlin et Bennett (2018)

considéré. Mais, au niveau de la dénotation primitive, la méthode tarskienne reste strictement énumérative, prenant la forme d'une longue liste de termes reliés à leurs référents. Si cela suffit pour construire une définition de la vérité extensionnellement correcte, cela est très loin aux yeux de Field de remplir des standards de réduction satisfaisants d'un point de vue physicaliste. Un physicaliste ne saurait se contenter d'une liste de noms associés à leurs dénotations et de prédicats associés à leurs extensions. Ce qu'il attend c'est une explication des faits physiques qui sous-tendent ces faits sémantiques [220]. Faute d'une telle explication, tout ce que donne la définition tarskienne c'est uniquement une réduction de la notion de vérité à d'autres notions sémantiques, à savoir la dénotation primitive des noms et des prédicats. Dès lors, Field suggère qu'une théorie de la vérité satisfaisante d'un point de vue physicaliste pourrait procéder en deux étapes : premièrement, elle s'appuierait sur les travaux de Tarski pour réduire la vérité à la dénotation primitive ; puis elle fournirait une théorie physicalistiquement acceptable de cette dernière notion. Autrement dit, la théorie tarskienne doit être complétée par une théorie physicalistiquement acceptable de la dénotation primitive expliquant à partir de seuls faits physiques comment les termes primitifs du langage acquièrent leurs valeurs sémantiques. Sur cette seconde étape, Field (1972) reste largement programmatique. Toutefois, il semble fonder de grands espoirs sur les théories causales de la référence récemment introduites à l'époque par Kripke [221]. Selon Field, en s'appuyant sur un réseau de lois causales reliant les termes du langage à leurs « actes de baptême », on devrait pouvoir obtenir une théorie satisfaisante de la dénotation primitive. Une fois ainsi complétée, la théorie de la vérité obtenue serait une forme de théorie correspondantiste compatible avec les exigences d'une bonne méthodologie scientifique [222]. Pour

220. Pour argumenter son propos, Field (1972) s'appuie sur une analogie avec la notion chimique de valence dont on aurait proposé une définition consistant simplement à lister chaque paire d'élément chimique associé à la valeur numérique de sa valence plutôt que d'en découvrir une véritable réduction expliquée à partir des configurations physiques de atomes. Nous ne rentrerons pas plus dans les détails de l'argumentation de Field ici. Pour plus de précisions nous renvoyons à Field (1972).

221. Voir Kripke (1972).

222. Bien entendu, cette théorie de la vérité correspondance aurait une forme peu usuelle. Rétrospectivement, Field lui-même la qualifie ainsi :

> Le format proposé pour une théorie de la correspondance dans [Field (1978, 1972)] était quelque peu non traditionnel : au lieu de parler explicitement d'une correspondance entre des énoncés (ou des états de pensée) et des faits ou des états de choses, j'ai parlé d'une correspondance (« référence primitive ») entre les composants des énoncés (ou des états de pensée) et les objets, les propriétés et autres choses du même genre. Mais [...] je pense que la conception défendue dans ces articles saisit l'idée principale derrière les théories de la correspondance plus traditionnelles. Ces deux articles considéraient la relation de correspondance comme une relation naturelle [N.D.T. *naturalistic relation*] dont il faut

être complet sur les positions du Field correspondantiste, signalons que l'article Field (1972) n'est pas le seul de cette période dans lequel Field tente de mettre au point et de défendre une théorie correspondantiste de la vérité acceptable pour un physicaliste. Vers la même époque (Field (1974, 1973)), il a également tenté de répondre au défi lancé par Quine et ses arguments d'indétermination [223]. De même, Field (1978) illustre comment pour Field les notions sémantiques, une fois dûment fondées, peuvent être employées en philosophie de l'esprit pour tenter de naturaliser les explications intentionnelles. Selon Field, une théorie de la vérité physicalistiquement acceptable peut aider à résoudre ce qu'il appelle le problème de Brentano, à savoir donner une explication compatible avec le matérialisme (ou le physicalisme) des attributions d'attitudes propositionnelles telles que la croyance que p, le désir que p, *etc.* , alors même qu'à première vue il s'agit là de propriétés « mentales » semblant relier des individus à des entités intentionnelles non matérielles, généralement appelées propositions [224]. On voit donc que le Field des années 1970 se trouvait au cœur d'un vaste programme de développement et fondation d'une théorie correspondantiste de la vérité acceptable d'un point de vue physicaliste, qui justifierait l'emploi de cette notion (et autres notions sémantiques) dans nos explications scientifiquement légitimes et autoriserait l'emploi essentiel des conditions de vérité dans l'analyse et la réduction des notions de signification (d'un énoncé) ou de contenu (d'un état mental représentationnel).

Au cours des décennies suivantes les conceptions de Field vont cependant connaître

fournir une explication théorique. (Field, 2001b, Préface p. vii)

223.

> [Les articles Field (1974, 1973)] furent conçus originellement comme faisant partie d'un programme visant à désamorcer la menace envers une théorie de la correspondance posée par les arguments d'indétermination de Quine (Field, 2001b, Préface p. viii)

Nous n'en dirons pas plus ici sur cet aspect du travail du jeune Field correspondantiste.

224. L'idée générale est la suivante : le fait pour un sujet S d'avoir une croyance que p est analysé comme le produit de deux relations, à savoir

(C) S croit que p si et seulement s'il existe une représentation mentale X telle que

(a) S *croit** X, et

(b) X *signifie* que p

où *croit** est une relation entre un sujet et un certain état mental, physiquement réalisé dans son cerveau et identifié au moyen d'une théorie computationnelle de l'esprit, relation qui ne pose aucun problème particulier pour un matérialiste ; tandis que la seconde relation —celle de *signification*— peut, selon Field, être expliquée de manière satisfaisante aux yeux d'un matérialiste (*i.e.* d'un physicaliste) en ayant recours à une théorie de la correspondance telle que celle esquissée dans Field (1972). Voyez Field (1978, en particulier les pages 40-43) —la pagination renvoie à Field (2001b)— pour une exposition détaillée des conceptions du jeune Field sur cette question.

un bouleversement fondamental. Pour comprendre ce changement, il faut se souvenir de l'importance du paradigme physicaliste au yeux de Field et des menaces que font peser sur lui les notions sémantiques. Le physicalisme semble intenable si les deux conditions suivantes sont conjointement vraies [225] :

1. Les notions sémantiques jouent un rôle dans les explications scientifiques.
2. Ces notions ne sont pas réductibles à des propriétés physiques.

Le renversement des positions de Field correspond à une inversion dans l'appréciation de la vérité respective de ces deux conditions. Le Field des années 1970 ne doutait pas que la condition 1. fût vraie et s'efforçait donc, nous venons de le voir, de montrer que la condition 2. était fausse en établissant une réduction ou, à tout le moins, une théorie physicalistiquement satisfaisante de la vérité correspondance. Au cours des années suivantes un glissement s'opère pour aboutir à un retournement complet au début des années 1990. D'une part, Field ne croit plus dorénavant à la promesse de la théorie causale de la référence [226], ni à la possibilité de mettre au point une théorie de la dénotation primitive permettant de réduire les explications sémantiques à des explications physiques. Autrement dit, Field doute désormais de la possibilité d'infirmer la condition 2. Mais, d'autre part, la vérité de la condition 1. ne lui semble plus aller de soi, et Field n'est plus persuadé à présent qu'une réduction des notions sémantiques soit réellement nécessaire. Les lectures de Quine (1970), Grover, Camp et Belnap (1975) et plus particulièrement de Leeds (1978) semblent avoir été déterminantes dans cette évolution [227]. Leeds (1978) insiste sur le fait que le paradigme physicaliste exige bien que les propriétés auxquelles un rôle *causal* est attribué dans nos explications soient expliquées en termes physiques, mais que cette exigence ne porte précisément *que* sur ces propriétés là. Pour les propriétés ou les notions n'entrant pas dans des relations causales, l'exigence de réduction ne porte pas. Or, Field en est de plus en plus persuadé, les notions sémantiques apparaissant dans nos explications n'y jouent qu'un rôle bien particulier : celui d'outils expressifs permettant de formuler des généralisations. Dans le Postscript ajouté à son article de 1972 repris dans Field (2001b, chapitre 1), Field décrit sa propre prise de conscience de ce point de la manière suivante :

> [...] l'affirmation selon laquelle il nous faut fournir une réduction physicaliste,

225. Nous empruntons cette élégante formulation à Galinon (2010, p. 53).
226. Force est en effet de constater que malgré les efforts déployés sur plus de trois décennies, une théorie causale de la référence réellement réalisable et qui fasse consensus semble encore hors de portée.
227. *Cf.* Field (2001b, p. 28-29, Postscript du chapitre 1).

> même approchante, de la notion de vérité, ou de conditions de vérité, ne tient que sous l'hypothèse que la vérité, ou les conditions de vérité, possèdent une sorte de rôle « causalement explicatif » [N.D.T. « *causal explanatory* » role]. (Ce point fut souligné dans Leeds (1978) et dans Putnam (1978)). Je pense que j'étais vaguement conscient de cela dans l'article, mais seulement vaguement. (Field, 2001b, p. 29, Postscript au chapitre 1)

Et quelque lignes plus loin :

> Je pense à présent que quoique ce rôle [central attribué à la vérité] soit sans nul doute de grande importance, il peut être expliqué entièrement à partir du rôle que joue la vérité en tant qu'outil de généralisation. [...] Si tel est le cas alors l'argument de ce chapitre en faveur d'une réduction physicaliste de la vérité était erroné. (Field, 2001b, p. 29, Postscript au chapitre 1)

Field se rapproche donc peu à peu d'une position très proche en esprit du décitationnisme de Quine. Il s'agira désormais non plus de fonder physiquement les propriétés sémantiques, mais de montrer qu'elles sont « causalement inertes » et ne jouent dans nos théories du monde qu'un rôle d'auxiliaires expressifs dénués de contenu explicatif propre. On passe ainsi d'une exigence de *réduction* à une exigence de *déflation* des notions sémantiques.

Dans Field (1994a,b), Field se déclare pour la première fois ouvertement déflationniste et développe sa nouvelle position. Field (1994a) distingue deux grandes traditions en philosophie du langage et en philosophie de l'esprit. Ces deux traditions

> diffèrent quant au rôle que la notion de conditions de vérité joue dans la théorie de la signification et dans la théorie du contenu des états intentionnnels. (Field (1994a), p. 104, la pagination renvoie à (Field, 2001b))

Une première tradition qu'il baptise tradition « inflationniste [228] » considère que

> les conditions de vérité jouent un rôle extrêmement central en sémantique et dans la théorie de l'esprit ; une théorie de la signification ou du contenu est, au moins pour une large part, une théorie des conditions de vérité. (Field, 1994a, p. 104)

C'est évidemment à cette première tradition que se rattachait le Field des années 1970 [229].

228. *Cf.* Field (1994a, p. 107).
229. Selon Field (1994a, p. 104) Frege, Russell, Ramsey et le Wittgenstein du *Tractatus* se classent également dans les rangs « inflationnistes ».

Field oppose à cette première tradition une seconde tradition qu'il nomme « déflationniste » et qu'il présente de la manière suivante :

> [...] l'idée principale derrière le déflationnisme [...] requiert seulement que ce qui joue un rôle dans la signification et le contenu ne comprenne pas les conditions de vérité (ou des relations envers les propositions, si les propositions sont conçues comme comprenant les conditions de vérité). (Field, 1994a, p. 108)

Field développe et approfondit cette idée quelques lignes plus loin et déclare :

> Si le déflationnisme est censé être un tant soit peu intéressant, il doit soutenir non seulement que ce qui joue un rôle central dans la signification et le contenu ne comprend pas les conditions de vérité *sous cette description*, mais qu'il ne comprend pas non plus *quoi que ce soit qui pourrait plausiblement constituer une réduction des* conditions de vérité à d'autres termes plus physicalistes. [...] En d'autres termes, une théorie du contenu et de la signification pourrait s'avérer ne pas employer la notion de conditions de vérité directement (dans un rôle central), mais employer (dans ce rôle) certaines relations physicalistes qui *pourraient* être regardées comme une réduction de la relation « S a les conditions de vérité que p » [...].
>
> (Field, 1994a, p. 108)

Autrement dit, le déflationniste se caractérise par son refus d'employer les conditions de vérité dans sa théorie du contenu ou de la signification, mais il doit en outre se garder de les employer subrepticement, et pour ainsi dire sans les nommer, en ayant recours à d'autres notions qui pourraient en constituer une réduction physicaliste. Après avoir ainsi délimité ces deux traditions opposées en philosophie du langage et de l'esprit, Field déclare qu'il

> ressent fortement les attraits de chacune des deux positions, bien qu'[il] en soit venu à préférer le déflationnisme.
>
> (Field, 1994a, p. 107)

À ce stade, ce nouveau parti pris théorique de Field en faveur du déflationnisme soulève de nombreuses questions. Premièrement, si les notions de vérité ou de conditions de vérité ne jouent pas de rôle central dans l'explication des concepts de signification et de contenu, comment ceux-ci doivent-ils être analysés ? Faut-il bannir purement et

simplement ces concepts de nos discours scientifiquement légitimes, ou bien, s'il faut les conserver, sur quelles autres notions —à l'exclusion des notions de vérité et apparentées— convient-il de s'appuyer pour en fournir un compte-rendu théorique ? Deuxièmement, si la notion de vérité n'intervient pas centralement dans l'analyse de la signification des énoncés et du contenu des états intentionnels, à quoi sert-elle ? Y a-t-il d'autres rôles qu'elle doive éventuellement remplir ou le discours sémantique n'est-il qu'un ornement dont il faudra se débarrasser en dernière analyse ? Enfin, une fois fournies les réponses à ces deux premières questions et une fois suffisamment clarifiée la conception déflationniste qui s'en dégage, quelle théorie de la vérité permettra d'articuler précisément ces intuitions ? À quoi doit ressembler, selon Field, une caractérisation précise d'un concept déflationniste de vérité ?

Commençons par la première question. À ce sujet, remarquons avec Field lui-même que le titre de son article[230] peut être source de confusion. En effet, la conception déflationniste qu'il entend désormais défendre

> pourrait être naturellement appelée *une conception déflationniste de la signification et du contenu* ; ou plus précisément, une conception déflationniste de *signifier que* et d'*avoir pour contenu que*, ou une conception déflationniste du *rôle des conditions de vérité dans la signification et le contenu*. La première de ces trois appellations [...] pourrait être trompeuse dans la mesure où certaines versions de la conception sont en un sens tout à fait non déflationnistes à propos de la signification.
>
> (Field, 1994a, p. 107, italiques de l'auteur)

Ainsi, contrairement à ce que pourrait laisser penser le titre de son article, Field n'entend pas « dégonfler » les notions de signification ou de contenu. Ce à propos de quoi Field est désormais déflationniste concerne uniquement le rôle de la vérité dans l'analyse de ces deux notions. La signification et le contenu peuvent tout à fait jouer eux-mêmes un rôle éminent, essentiel et crucial dans notre entreprise théorique. Simplement, l'analyse de ces deux notions devra s'effectuer sans faire appel essentiellement à la notion de vérité et autres propriétés sémantiques.

Comme première illustration élémentaire d'une analyse déflationniste en ce sens, Field (1994a) mentionne l'exemple de la théorie vérificationniste de la signification.

230. En version originale, ce titre est *Deflationist Views of Meaning and Content* ; ce qu'on pourrait traduire par *Les conceptions deflationnistes de la signification et du contenu.*

Grosso modo, selon ce paradigme l'analyse de la notion de signification d'un énoncé doit se faire à partir de ses conditions de vérification et non pas en s'appuyant sur ses conditions de vérité. La notion de signification est donc expurgée de la notion de conditions de vérité [231], mais cela ne l'empêche pas d'être à son tour un pilier théorique fondamental de l'explication (vérificationniste) du langage et de l'esprit. Ceci étant, Field lui-même n'entend pas se limiter à un vérificationnisme rudimentaire et il donne dans son article de nombreux indices sur ce que pourraient être à ses yeux les ingrédients acceptables d'une analyse déflationniste de la signification et du contenu —déflationniste au sens qui vient d'être précisé. Selon Field (1994a), pour analyser la signification d'un énoncé ou le contenu d'un état mental, un déflationniste pourra s'appuyer sur la sémantique des rôles conceptuels [232] :

> Un élément qui peut certainement être inclus dans le contenu est le rôle conceptuel ou computationnel : le rôle dans une psychologie computationnelle (peut-être idéalisée) qui décrit comment les croyances, les désirs, *etc.* de l'agent évoluent au cours du temps (notamment en réponse aux stimulations sensorielles). Le rôle conceptuel d'un état de croyance inclut ses conditions de vérification, mais il inclut bien d'autres choses en plus.
>
> (Field, 1994a, p. 108-109)

À ces premiers éléments s'ajoutent ce que Field appelle les « relations d'indications » et qu'il décrit comme de simples corrélations statistiques entre l'apparition de certains états de croyance chez un agent et l'occurrence de certains événements dans son environnement.

> Les relations d'indication sont [un] élément [...] qu'un déflationniste peut intégrer dans le contenu. C'est un fait me concernant que je suis un assez bon baromètre pour détecter s'il tombe de la pluie sur ma tête en ce moment : quant il y a de la pluie tombant sur ma tête, j'ai tendance à croire « De la pluie tombe sur ma tête » ; à l'inverse, quand je crois effectivement cet énoncé, en général il y a de la pluie tombant sur ma tête. *Il s'agit simplement là d'une corrélation, que l'on peut observer* ; et un déflationniste est aussi libre que

231. Du moins au sens traditionnel, « inflationniste » que possède habituellement ce terme. Voyez Field (1994a, p 104-108) pour une discussion plus approfondie.

232. Sur la sémantique des rôles conceptuels et son rapport avec la sémantique vériconditionnelle, voyez Harman (1982) et Loar (1982). Pour un bilan plus récent, voir Greenberg et Harman (2008) et Whiting (2018).

> n'importe qui de la remarquer et aussi libre que n'importe qui de la considérer comme un ingrédient de ce qu'il appelle le contenu.
>
> (Field, 1994a, p. 109, nous soulignons)

Puis Field ajoute :

> La corrélation des états de croyance des gens avec le monde qui les entoure s'étend probablement au delà de ce qui est directement observable.
>
> (Field, 1994a, p. 109)

Mais en incorporant dans notre analyse du contenu et de la signification les notions de rôle conceptuel et les relations d'indication, ne court-on pas le risque de succomber au danger déjà évoqué de reconstruire subrepticement une notion forte de conditions de vérité qui viendrait clandestinement s'insérer dans notre explication de la signification ou du contenu ? Ce point est crucial pour la démarche défendue par Field. Le pari de Field est que l'on peut répondre à cette question par la négative !

Il déclare tout d'abord :

> Ces observations pourrait faire penser qu'un déflationniste est condamné à reconnaître (une version non-décitationnelle de) la relation « S a pour conditions de vérité que p », dans les faits si ce n'est dans les termes [N.D.T *in fact if not in name*], *puisque ces relations d'indication constituent la relation de conditions de vérité*. Mais c'est négliger le fait que le projet de fournir quoi que ce soit qui ressemble à une réduction crédible du discours en termes de conditions de vérité à un discours en termes de relations d'indication est tout au plus une lueur dans l'œil de certains théoriciens.
>
> (Field, 1994a, p. 110, nous soulignons)

Pour illustrer cette distinction, Field donne des exemples d'énoncés dont les relations d'indication, *i.e.* leurs correlations statistiques avec certains événements apparaissant dans l'environnement du locuteur, semblent diverger largement de leurs conditions de vérité [233]. Field indique en outre qu'

> il n'est pas inimaginable que même certains de nos propres compte-rendus observationnels indiquent de manière fiable quelque chose qui diffère de leurs propres conditions de vérité.
>
> (Field, 1994a, p. 110)

233. Nous ne rentrerons pas dans les détails ici. Nous renvoyons le lecteur à Field (1994a, p. 110-112).

Au final, quelques pages plus loin, Field conclut de la manière suivante sur ce problème de la possibilité de reconstruire, dans les faits si ce n'est dans les termes, une relation inflationniste de conditions de vérité :

> Mon hypothèse est que cela se révélera ne pas être le cas.
>
> (Field, 1994a, p. 119)

Toutefois Field a bien conscience qu'il s'agit là d'un pari et qu'on ne peut être certains à l'avance du résultat final de nos investigations. C'est pourquoi il milite pour une forme de « déflationnisme méthodologique » :

> [N]ous devrions être des « déflationnistes méthodologiques » : c'est-à-dire que nous devrions démarrer en postulant le déflationnisme comme hypothèse de travail ; nous devrions l'accepter à moins que et jusqu'à ce que nous nous retrouvions à reconstruire ce qui équivaudrait à la relation inflationniste « S a pour conditions de vérité que p ».
>
> (Field, 1994a, p. 119)

Voici donc en quoi consiste la gageure du déflationnisme méthodologique : faire l'hypothèse par défaut que les notions sémantiques ne jouent pas de rôle central dans l'analyse de la signification et du contenu et pousser cette hypothèse aussi loin que possible jusqu'à réalisation complète d'un programme déflationniste ou jusqu'à ce que le recours une notion « robuste », « inflationniste » de conditions de vérité s'avère finalement indispensable dans nos explications.

Passons donc à la seconde question : si les notions sémantiques n'interviennent plus dans notre théorie du contenu ou de la signification, à quoi peuvent-elles bien servir ? La réponse de Field à cette question n'est guère surprenante et se situe dans la droite ligne des réflexions déflationnistes contemporaines. Nous serons donc brefs sur ce point. À l'instar de Quine (1970) ou d' Horwich (1998b), Field rappelle que pour le déflationniste,

> le mot « vrai » possède un important rôle logique : il nous permet de formuler certaines conjonctions et disjonctions infinies que nous ne pourrions pas formuler autrement.
>
> (Field, 1994a, p. 120)

Bien qu'elle soit rigoureusement similaire à celles que nous avons déjà rencontrées avec Quine et Horwich, nous donnons une nouvelle illustration de ce phénomène, due cette fois à Field :

> « Réalisme » a été employé pour signifier bien des choses, mais une de ses versions est la thèse selon laquelle il se pourrait qu'il existe (et qu'il existe très certainement) des énoncés de notre langage qui sont vrais mais que nous n'aurons jamais de raisons de croire [...]. Pour proclamer le réalisme en ce sens, il nous faut une notion de vérité. Mais la raison de cela est purement logique : c'est-à-dire que si on ne pouvait formuler qu'un nombre fini d'énoncés dans notre langage, nous pourrions exprimer la doctrine réaliste sans employer de prédicat de vérité : nous pourrions dire
>
> > Il se pourrait que le nombre de brontosaures qui ont jamais vécu soit précisément 75 278 mais que nous n'aurons jamais de raison de le croire ; ou que le montant total dépensé par Michael Jackson pour ses sous-vêtements au cours de sa vie soit exactement de 1 078 852,72 $ mais que nous n'aurons jamais de raison de le croire ; ou ...
>
> où à la place des « ... » se trouveraient des clauses similaires pour chaque énoncé du langage. (Field, 1994a, p. 120)

Autrement dit, pour Field, comme pour Quine ou Horwich, ce qui rend le prédicat de vérité indispensable c'est son rôle en tant qu'outil [*device*] expressif permettant de formuler des généralisations telles que des conjonctions et des disjonctions infinies.

Nous avons donc à présent une explication assez claire de ce à quoi peut servir un prédicat de vérité dans le cadre du déflationnisme méthodologique de Field. Avec la réponse à la question précédente, nous avons également une idée assez claire de ce à quoi il *n'est pas* censé servir. Nous pouvons donc maintenant nous tourner vers notre troisième question. Comment mettre en forme plus précisément les conceptions en matière de vérité que nous venons de décrire ? Autrement dit, quelle forme peut prendre une théorie d'un prédicat de vérité déflationniste au sens fieldien du terme ? Ce point soulève quelques difficultés. En effet, Field lui-même oscille entre plusieurs formulations possibles pour sa théorie de la vérité, dont il n'est pas certain qu'elles soient toutes en harmonie les unes avec les autres. Dans le seul article Field (1994a), nous avons pu relever pas moins de cinq formulations différentes avancées par Field [234] :

234. Gupta et Martínez-Fernández (2005, p. 56, note 11) remarquent eux aussi qu'en l'espace d'une seule page imprimée Field (1994a, p. 114-115) propose, pour axiomatiser son prédicat de vérité, pas moins de trois formulations différentes, en plus de celle indiquée en début d'article. Signalons également qu'*en plus des cinq formulations que nous citons ici*, Field introduit encore d'autres variantes d'axio-

1. Une première formulation (Field, 1994a, p. 105 et p. 121) postule que, sous l'hypothèse que l'énonciation [*utterance*] u existe,

   L'affirmation que u est vraie est cognitivement équivalente à u elle-même.

2. Plus loin, Field (1994a, p. 114) met cette fois-ci en avant le « schéma de décitation » pour un énoncé [*sentence*] p :

   (T) « p » est vrai si et seulement si p

   dont il précise qu'il est une « nécessité conceptuelle ».

3. Quelques lignes plus bas, Field (1994a, p. 115) propose de choisir plutôt comme point de départ une version généralisée de (T) qui emploie la quantification substitutionnelle :

   (TG) Πp (« p » est vrai si et seulement si p)

4. Trois lignes après, Field (1994a, p. 115) suggère encore une autre alternative qui consiste à employer des lettres schématiques pour les énoncés et à raisonner avec elles comme avec des variables au moyen des règles d'inférence suivantes :

   (i) Toute instance d'une lettre schématique peut être remplacée par un énoncé.

   (ii) à partir du schéma $A($« p »$)$ dans lequel toutes les occurrences de p figurent entre guillemets, on peut déduite $\forall x(\text{Énoncé}(x) \rightarrow A(x))$.

5. Enfin, Field (1994a, p. 123) affirme que dans le cas où notre langage contient des opérateurs modaux, il nous faut une axiomatisation un peu plus forte et que le schéma (T) doit alors être remplacé par

   (NT) $\Box$ (« p » est vrai si et seulement si p).

Nous ne prétendrons pas étudier en détails les mérites comparés et la compatibilité de ces diverses formulations [235]. Mais quoi qu'il en soit de ces variations, Field présente systématiquement son prédicat de vérité comme un prédicat décitationnel. Afin d'examiner

matisation pour traiter ce qu'il appelle la notion quasi-décitationnelle de vérité (voyez Field (1994a, p. 131, et le Postscript).)

235. Remarquons simplement en passant que la formulation 3. ci-dessus, qui emploie la quantification substitutionnelle, a de quoi étonner. Sur ce point, Field se détache de Quine et d'Horwich qui, nous l'avons vu, rejettent ce type de quantification. De plus, Field affirme que c'est le rôle du prédicat de vérité en tant qu'outil expressif pour formuler des conjonctions et des disjonctions infinies qui explique son indispensabilité et sa présence dans nos ressources théoriques. Or, nous avons déjà vu —voyez la discussion de ce point dans la section sur Ramsey— que si l'on accepte la quantification substitutionnelle, un tel rôle n'a plus lieu d'être. Le prédicat de vérité devient dès lors explicitement définissable et éliminable. Field lui-même cite ce résultat dans une note (Field, 1994a, p. 120, note 17) mais ne semble pas le considérer comme problématique pour la formulation 3. et ses variantes.

certaines caractéristiques de la vérité décitationnelle telle qu'elle est comprise par Field, nous retiendrons donc ici la caractérisation générale suivante tirée d'un autre article paru à la même époque :

> Le déflationnisme est la thèse selon laquelle la vérité est fondamentalement [N.D.T. *at bottom*] décitationnelle. Je considère que cela signifie que dans son usage primaire (« purement décitationnel »),
> — « vrai » en tant qu'il est compris par une personne ne s'applique qu'aux énonciations [N.D.T *utterance*] que cette personne comprend, et
> — pour toute énonciation u qu'une personne X comprend, l'assertion que u est vraie est cognitivement équivalente pour X à u elle-même.
> (Field (1994b), p. 222, la pagination renvoie à (Field, 2001b))

Cette spécification d'un prédicat de vérité décitationnel emploie les notions de compréhension (d'une énonciation par une personne) et d'équivalence cognitive (entre deux énonciations ou deux assertions). Dans Field (1994a,b), Field ne donne guère d'indications sur la notion de compréhension à laquelle il est fait référence ici et renvoie vraisemblablement à un usage courant et intuitif de ce concept. Remarquons toutefois que puisque cette notion apparaît dans la caractérisation du prédicat « vrai », elle doit sans doute être analysée antérieurement et indépendamment de la propriété de vérité (faute de quoi, on succomberait à un cercle vicieux). Pour le dire dans une formulation plus fieldienne, les conditions de vérité ne peuvent visiblement pas jouer de rôle central dans l'explication de la notion de compréhension. Par conséquent, comprendre un énoncé ne peut pas consister fondamentalement à saisir ses conditions de vérité. Ce point peut sembler contre-intuitif et s'oppose en tout cas à une large tradition classique en philosophie du langage selon laquelle comprendre (la signification d') un énoncé, c'est précisément savoir à quelles conditions il est vrai. On peut néanmoins faire crédit à Field ici d'une certaine cohérence : après tout, si par exemple comprendre un énoncé, c'est en saisir la signification et si la signification doit être analysée sans recourir à une notion de conditions de vérité, on peut s'attendre à ce que ces dernières ne jouent pas non plus de rôle dans l'explication de la compréhension. Dans le même ordre d'idée, on peut

À l'inverse dans leur commentaire de l'article de Field, Gupta et Martínez-Fernández considèrent que cette formulation

> peut difficilement compter comme une version du déflationnisme puisqu'elle n'attribue à la vérité aucun rôle logique essentiel. Elle rend la vérité redondante. (Gupta et Martínez-Fernández, 2005, p. 56, note 11)

imaginer que comprendre un énoncé s'analyse plutôt comme la capacité d'un locuteur à en saisir le rôle conceptuel, les relations d'indication et les éventuels autres ingrédients que Field accepte pour l'analyse du contenu et de la signification. Même si Field reste muet sur ce sujet précis, on peut considérer qu'il s'agit d'une conséquence naturelle de son déflationnisme méthodologique.

À l'inverse, Field fournit quelques précisions sur la manière dont il faut comprendre la notion d'équivalence cognitive. Dans une note de bas de page, il écrit :

> Je considère l'équivalence cognitive comme étant une question de rôle conceptuel ou computationnel : pour un énoncé, être cognitivement équivalent à un autre pour une personne donnée, c'est être tel que les règles d'inférence de cette personne permettent (ou permettent de manière relativement directe) l'inférence de l'un à l'autre et réciproquement. (Field, 1994b, p. 222, note 1)

Là encore, l'équivalence cognitive ne peut pas s'analyser en termes d'identité ou de préservation de conditions de vérité de chacun de ses membres à la manière d'une (certaine conception de l') équivalence logique. De même, « permettre l'inférence » ne peut bien sûr pas s'expliquer en recourant à une notion de préservation de la vérité, car la notion d'équivalence cognitive est censée être plus fondamentale et participer à la caractérisation même de la notion de vérité. Field suggère donc une approche inférentialiste de l'équivalence cognitive : elle doit être analysée à partir des notions de rôle conceptuel et computationnel, et sans qu'interviennent de termes sémantiques.

Mais même une fois clarifiées les ressources sur lesquelles Field s'appuie pour mettre au point sa théorie de la vérité, il reste que sa conception « purement décitationnelle » du prédicat « vrai » renferme des conséquences surprenantes. Une première série de conséquences a trait au fait que le prédicat de vérité, tel que caractérisé par Field, ne s'applique, pour un locuteur donné, qu'aux énoncés que ce locuteur comprend ; une seconde série de conséquences concerne le statut modal de l'équivalence cognitive reliant un énoncé p à l'énoncé « p est vrai ». Ces conséquences peuvent paraître déroutantes et s'éloignent, voire s'opposent, à certains emplois habituels de la notion de vérité. Nous allons en dire quelques mots, mais notons dès à présent que Field a pleinement conscience de ces conséquences contre-intuitives et qu'il les assume ouvertement. Bien plus, il va jusqu'à affirmer qu'

> [e]n tout cas, ces deux caractéristiques de la vérité purement décitationnelle

> la rende idéalement adaptée pour remplir le besoin logique d'un outil de conjonction et de disjonction infinies [...] (Field, 1994a, p. 122)

Il faut se souvenir que le projet de Field ne consiste pas tant à théoriser ou expliquer notre notion naïve et pré-théorique de vérité[236] qu'à développer un concept de vérité qui remplisse les objectifs théoriques qu'il s'est fixé dans le cadre du déflationnisme méthodologique. Ces objectifs, rappelons le, cantonnent les notions sémantiques à être de simples outils expressifs répondant à un besoin logique précis —celui de formuler des disjonctions et des conjonctions infinies— mais sans contenu explicatif propre, ce qui permet de justifier leur emploi, sans réduction, dans un cadre physicaliste. Field ne cache d'ailleurs pas son intérêt limité envers les emplois du prédicat « vrai » dans le langage courant :

> Cependant, je me méfie de ce genre d'affirmations concernant ce que les personnes ordinaires veulent dire lorsqu'elles forment des assertions contenant « vrai » : je doute plutôt qu'il existe une manière cohérente de rendre compte de tous les emplois ordinaires de cette notion. J'ai plutôt tendance à penser que de *nombreux* emplois ordinaires de « vrai » rentrent effectivement dans le moule purement décitationnel, quoique je regarde la question de savoir dans quelle mesure c'est le cas comme d'un intérêt uniquement sociologique. (Field, 1994a, p. 133, italiques de l'auteur)

Revenons à notre première série de conséquences. Le fait que le prédicat de vérité purement décitationnel ne s'applique qu'aux énoncés de ce que Field appelle parfois l'« idiolecte » propre à un individu donné, c'est-à-dire l'ensemble des énoncés que cet individu comprend à un temps donné, l'éloigne évidemment de la notion habituelle de vérité. En effet, je ne crois pas commettre d'incorrection lorsqu'en tant que locuteur francophone compétent j'affirme : « il existe sans doute des énoncés de volapük qui sont vrais », quand bien même je ne comprendrais pas un mot de volapük. De même, dans la mesure où l'ensemble des énoncés compris varie d'un locuteur à l'autre, ainsi qu'au cours du temps pour un même individu fixé, l'extension du prédicat purement décitationnel caractérisé par Field va différer selon l'individu considéré et fluctuer au gré des apprentissages et des oublis modifiant l'idiolecte d'un individu. C'est une conséquence que Field (1994a) accepte sans sourciller. Il affirme même que

236. Si tant est qu'une telle unique notion correspondant aux emplois de la vérité dans le langage courant existe.

> les seuls énoncés avec lesquels nous formons jamais littéralement des conjonctions ou des disjonctions sont des énoncés que nous comprenons, il est donc clair qu'une notion inapplicable aux énonciations que nous ne comprenons pas servira nos besoins de vérité en tant qu'outil de conjonction et de disjonction. (Field, 1994a, p. 122).

Malgré tout, Field ne renonce pas à la possibilité de rendre compte des attributions de vérité à des énoncés d'une langue étrangère ou d'un langage incompris à partir de la notion purement décitationnelle de vérité mais *associée à d'autres ressources*. Examiner en détails ses propositions nous mènerait toutefois trop loin et nous renverrons donc le lecteur à Field (1994a, p. 127-130, section 8) et à Field (2001b, p. 147-151, Postscript au chapitre 4) [237].

Le second type de conséquences de la conception purement décitationnelle développée par Field, qui concerne le statut modal des équivalences cognitives, est encore plus déconcertant. Tout d'abord, pour Field, c'est l'équivalence *cognitive* entre les énoncés p et les énoncés « « p » est vrai (au sens purement décitationnel) » qui *fonde* les équivalences décitationnelles classiques du décitationnisme :

237. Très rapidement : Field (1994a, p. 127-130) explore essentiellement trois routes pour expliquer les attributions de vérité à un énoncé d'une langue étrangère.

1. La première possibilité consiste à employer une notion primitive de *synonymie interlinguistique*

   « S est vrai » (où « S » est un énoncé d'une langue étrangère) est équivalent à « S est *synonyme* d'un énoncé de mon idiolecte vrai au sens purement décitationnel ».

   Néanmoins Field doute que la notion de synonymie interlinguistique puisse s'intégrer à son déflationnisme méthodologique.

2. La vérité d'un énoncé d'une langue étrangère peut aussi être expliquée relativement à une *corrélation* entre cette langue et mon propre idiolecte :

   « S est vrai relativement à telle corrélation » (où « S » est un énoncé d'une langue étrangère) est équivalent à « S est *corrélé* avec un énoncé de mon idiolecte vrai au sens purement décitationnel ».

   La notion de corrélation entre deux langues est semblable à celle de traduction à ceci près que ses standards de correction sont hautement relatifs aux contextes et aux intérêts poursuivis (en particulier aucune idée de préservation des conditions de vérité n'entre en jeu).

3. La dernière option consiste à appliquer sans relativisation le concept purement décitationnel aux énoncés d'une langue étrangère que je comprends (et uniquement à ceux-là). Mes connaissances en anglais me permettent ainsi d'affimer, malgré les entorses apparentes à la grammaire :

   « Snow is white » est vrai si et seulement si snow is white.

Enfin, dans le Postscript ajouté à Field (1994a) à l'occasion de sa re-parution dans Field (2001b), Field suggère que sa restriction aux énoncés compris pas un locuteur n'était pas la bonne. Dorénavant, on doit pouvoir appliquer le prédicat « vrai » à tout énoncé déclaratif e de quelque langue que ce soit. De par la thèse de l'équivalence cognitive, dans la mesure où on ne comprend pas e, on ne comprend pas non plus « e est vrai ». Mais les deux énoncés sont néanmoins liés par une équivalence. Il s'agit là d'un cas de ce que Field appelle « l'indétermination corrélée » (*cf.* Field (2001b, p. 147-151, Postscript)).

(T) « p » est vrai si et seulement si p

telles qu'on les retrouvent chez Quine (ou sous une forme propositionnelle chez Horwich). Et cela va avoir des conséquences importantes sur le statut modal et la nature de ces équivalences qui vont être bien plus que de simples équivalences matérielles. On retrouve cette idée à de nombreux endroits dans les écrits de Field. Dans Field (1994b), Field affirme ainsi que pour un locuteur X qui comprend le prédicat « vrai » dans son sens purement décitationnel, les instances du schéma de décitation (T) ci-dessus dans lesquelles on remplace les occurrences de « p » par un énoncé que X comprend, seront

> plus ou moins [N.D.T. *more or less*] « analytiques » ou « logiquement vraies » pour X, en vertu de l'équivalence cognitive des membres de gauche et de droite. (Field, 1994b, p. 222)

De même, Field (1994a, p. 114) déclare que

> chaque instance du « schéma de décitation »
>
> (T) « p » est vrai si et seulement si p
>
> est une nécessité conceptuelle [...] en vertu de l'équivalence cognitive des membres de gauche et de droite. (Field, 1994a, p. 114)

Field va même encore plus loin puisqu'il affirme que si le langage est muni d'un opérateur modal,

> le schéma (T) devrait être remplacé par
>
> (NT) □ (« p » est vrai si et seulement si p),

ce qui revient à dire que les énoncés possèdent leurs conditions de vérité *par nécessité*[238]. Ceci peut paraître extrêmement contre-intuitif, mais Field assume parfaitement cette conséquence de sa conception purement décitationnelle de la vérité. Selon lui, l'équivalence cognitive entre un énoncé et l'attribution de vérité au sens purement décitationnel à cet énoncé implique que

238. Si par exemple on interprète la nécessité modale □ en termes de mondes possibles, alors

□ (« la neige des blanche » est vrai si et seulement si la neige est blanche)

signifie que dans tout monde possible « la neige est blanche » a pour conditions de vérité que la neige est blanche, y compris dans un monde possible où les locuteurs francophones emploient le substantif « neige » pour désigner le charbon.

> la notion [de vérité décitationnelle pure] est une propriété *indépendante de l'usage* [N.D.T. use-independent *property*] : qualifier « la neige est blanche » de décitationnellement vrai, c'est simplement qualifier la neige de blanche ; ce *n'*est donc *pas* lui attribuer une propriété que [cet énoncé] n'aurait pas possédée si les autres locuteurs francophones et moi-même avions employé les mots d'une manière différente. (Field, 1994a, p. 122, italiques de l'auteur)

Aux yeux de Field, cette caractéristique paradoxale de la vérité décitationnelle est même une qualité puisqu'elle permet au prédicat « vrai » de remplir parfaitement le rôle d'outil expressif de généralisation que Field lui attribue :

> l'indépendance par rapport à l'usage de la vérité décitationnelle est *requise* par les objectifs que l'on vient de revoir [239]. Car si « Tous les énoncé de type Q sont vrais » doit servir comme une conjonction infinie de tous les énoncés de type Q, alors nous voulons qu'il implique chacun de ces énoncés, et qu'il soit impliqué par eux, pris tous ensemble. Ce ne serait pas le cas à moins que « S est vrai » n'implique et ne soit impliqué par S. Mais la seule manière pour qu'il en soit ainsi est que « vrai » n'attribue pas à S une caractéristique dépendante de l'usage. Supposons par exemple que la géométrie euclidienne soit vraie, et que nous tentions d'exprimer sa nature contingente en disant que l'ensemble de ses axiomes aurait pu être faux. Assurément, ce que nous voulions dire n'était pas simplement que les locuteurs auraient pu employer leurs mots de telle manière que les axiomes ne soient pas vrais, mais que l'espace lui-même aurait pu être différent de façon à rendre faux les axiomes *tels que nous les comprenons*. Une notion de vérité indépendante de l'usage est précisément ce qu'il nous faut. (Field, 1994a, p. 122, italiques de l'auteur)

Aussi bizarres que puissent paraître ces particularités de la vérité « purement décitationnelle », Field semble donc les accepter volontiers.

Toutefois sa pensée a connu une évolution sur ce point. Le problème est que la vérité purement décitationnelle ainsi caractérisée comme indépendante de l'usage, heurte certaines de nos intuitions modales, notamment dans les contextes de contrefactuels. Par exemple, comment rendre-compte d'un énoncé tel que

239. N.D.T Field fait ici référence au rôle du prédicat de vérité pour l'expression de conjonctions et de disjonctions infinies

« Si nous avions employé le terme « neige » de la manière dont nous employons le terme « charbon », alors l'énoncé « la neige est noire » aurait été vrai »

au moyen d'un prédicat de vérité purement décitationnel ? Déjà dans Field (1994a, p. 130-133, section 9), Field envisage d'introduire une autre notion de vérité qu'il appelle « quasi-décitationnelle » et qui serait plus fidèle à nos intuitions modales, en étant pour partie dépendante de l'usage. Cependant, cette autre notion de vérité

> requiert que nous possédions non seulement une notion de synonymie mais aussi une notion antérieure de signification telle que deux énoncés soient synonymes s'ils ont la même signification ; *mais la signification doit être définie indépendamment des conditions de vérité.* (Field, 1994a, p. 131, nous soulignons)

Avec ces ressources, le schéma de vérité (T) peut être remplacé par la caractérisation suivante :

> $\Box \Big($S est vrai$_{\text{quasi-décitationnellement}}$ ssi $\Sigma p[\exists m(m$ est la signification de S et $@(m$ est la signification de « p »)) et $p]\Big)$

> [...] où « @ » est un « opérateur de réalité [N.D.T *actually operator*] » qui « annule temporairement les effets de » l'opérateur modal. Ceci semble avoir les propriétés modales désirées si nous voulons imiter un prédicat de vérité inflationniste pour rendre la vérité des énoncés dépendantes de l'usage. (Field, 1994a, p. 131-132)

Bien entendu, l'introduction des notions de synonymie et de signification est un danger potentiel pour le déflationnisme méthodologique. Field affirme toutefois que cette conception quasi-décitationnelle peut encore être considérée comme

> quelque peu déflationniste [N.D.T. *somewhat deflationary*], si l'idée de signification est expliquée d'une manière suffisamment « non-vériconditionnelle » [N.D.T. *« un-truth-theoretic » way*]. (Field, 1994a, p. 132)

Quelques années plus tard, dans le Postscript ajouté à son article Field (1994a), Field revient encore sur ce problème et modifie sa position. Dorénavant, c'est la notion de traduction —expliquée en des termes suffisamment « non-vériconditionnels »— qui est mise en avant. Et Field affirme que plutôt que de distinguer une notion purement décitationnelle et une notion quasi-décitationnelle de vérité,

> nous pouvons utiliser un unique prédicat de vérité tant que nous prenons comme entités dans son extension non pas les types orthographiques mais les types computationnels : des classes d'équivalence de tokens (potentiels) pour la relation d'équivalence computationnelle. (Field, 2001b, p. 151, Postscript au chapitre 4)

Field élabore un peu plus sa notion d'équivalence computationnelle entre des énoncés pris comme types dans Field (2001a). Sans entrer dans les détails, l'idée générale est que les énoncés-types ne doivent plus être identifiés orthographiquement ou phonétiquement comme des suites de graphèmes ou de phonèmes, mais plutôt à partir de leurs rôles conceptuel dans la psychologie du locuteur qui les emploie. De ce point de vue, on peut les identifier à des classes formées par les énoncés-tokens tirés de l'idiolecte d'un agent que cet agent traite de manière équivalente sur un plan computationnel. Avec cette identification des énoncés à des types computationnels, on peut alors, selon Field, expliquer la vérité au moyen du schéma suivant :

> (***) Si $S_{X,u}$ est traduisible par « p » alors $\Box$($S_{X,u}$ est vrai ssi p)
>
> [Et si $S_{X,u}$ est traduisible par « p » alors $\Box$($S_{X,u}$ a pour conditions de vérité que p)]
> (Field, 2001b, p. 152, Postscript du chapitre 4)

où X est un individu donné, situé dans un monde possible u, et où $S_{X,u}$ est un énoncé identifié par le rôle conceptuel qu'il tient dans l'idiolecte du locuteur X dans le monde u ($S_{X,u}$ n'est donc pas identifié phonétiquement ou orthographiquement).

L'emploi des notions de traduction (ou de signification) et l'identification des porteurs de vérités à des types computationnels ne marquent-ils pas le retour des notions sémantiques « inflationnistes » dans notre théorie de la vérité ? Field semble confiant que non. À la fin du passage de son Postscript consacré à cette question, il écrit

> l'argument pour une théorie substantielle des conditions de vérité et de la référence dépend du fait qu'on considère les conditions de vérité et la référence comme jouant un certain type de rôle « causalement explicatif [N.D.T. *causal explanatory*] [...] ; et introduire une notion de conditions de vérité au moyen du schéma (***) ne fait rien pour rendre la notion « causalement explicative ».

Nous laisserons pour notre part cette question en suspend et renvoyons le lecteur intéressé par les réflexions et les inflexions des conceptions de cet auteur essentiel du déflationnisme contemporain à Field (2005a,b,c,d, 2001b).

1.3 Bilan

Au terme de cette excursion dans l'histoire du déflationnisme, nous avons pu mettre en lumière un certain nombre de thèses caractéristiques du déflationnisme contemporain. Comme nous l'avons constaté, les auteurs déflationnistes actuels estimment que la vérité n'est pas semblable aux propriétés ordinaires. Selon eux, le prédicat de vérité n'est qu'un outil de décitation (ou de dénominalisation) dont la seule raison d'être est de nous permettre de formuler certaines généralisations qui s'apparentent à des disjonctions ou des conjonctions infinies. Cette *unique* fonction du prédicat de vérité le rend indispensable. Dans ce type de conceptions, la vérité est entièrement caractérisée par la collections des **T**-équivalences , puisque cette axiomatisation, ou cette définition implicite [240], permet de rendre compte de tous les emplois décitationnels du prédicat « vrai ». En ne considérant la vérité que comme un simple auxiliaire expressif comblant un besoin logique, les déflationnistes suggèrent également que la vérité n'a pas de contenu propre. En tout état de cause, ils soutiennent qu'elle ne saurait jouer de rôle explicatif central dans nos théories.

Dans la partie qui suit immédiatement ce chapitre, nous allons en quelque sorte prendre les déflationnistes « au mot » lorsqu'il affirment que la vérité est une sorte de notion logique. Pour ce faire, nous nous appuierons sur un cadre méthodologique inspiré des caractérisations inférentielles de la logicité. Dans la partie suivante, nous examinerons un argument qui repose sur une autre manière de préciser les dires déflationnistes. Cette fois, les allégations de « non-substantialité », d'absence de contenu propre, ou de rôle explicatif, censées caractériser la vérité déflationnistes sont traduites par une propriété de conservativité.

Quoiqu'elles renferment des thèmes communs, notamment le rôle crucial qu'y joue la notion de conservativité, ces deux parties sont largement indépendantes [241]. Elles constituent deux tentatives complémentaires de donner un sens précis aux thèses des déflationnistes pour pouvoir les confronter dans un cadre méthodologique rigoureux à des résultats techniques ou formels précis afin d'en évaluer plus exactement la portée et la solidité.

240. en un sens relâché
241. On peut d'ailleurs les lire de manière autonome.

Première partie

Vérité et Logicité

Chapitre 2

Vérité et Logicité

2.1 Préambule : une thèse déflationniste sur le prédicat de vérité.

Dans le chapitre précédent, nous avons pu constater que derrière l'étendard du déflationnisme contemporain se regroupent divers auteurs, aux conceptions et thèses parfois variées. Un point commun à tous ces auteurs que l'on peut néanmoins dégager est l'idée selon laquelle le prédicat « vrai » serait avant tout un outil expressif et ne désignerait nullement une propriété ayant un réel contenu. La vérité ne serait pas une propriété « substantielle », dont les mystères resteraient à découvrir au moyen d'une subtile analyse philosophique.

Ainsi, Quine (1970, 1990) [1] analyse l'usage du prédicat comme un simple outil de décitation permettant d'annuler la montée sémantique. L'introduction du prédicat « vrai » permet d'annuler l'effet des guillemets et nous ramène vers le monde :

Asserter : « l'énoncé « la neige est blanche » est vrai »
revient à asserter : « la neige est blanche ».

Alors qu'il semble que nous attribuions une propriété à un énoncé placé entre guillemets, en réalité nous parlons de la blancheur de la neige. Plus récemment Horwich (1998b) affirme que tout ce qu'il y a à dire sur la vérité est donné par la théorie minimale qui consiste en la collection inifinie des instances du schéma suivant :

(Min) La proposition que p est vraie si et seulement si p.

1. Voyez le chapitre précédent pour des références plus précises.

Maîtriser le concept de vérité consiste simplement à être prêt à accepter toutes les instances de (Min), rien de plus. Par ailleurs, ajoute Horwich, la collection infinie des instances de (Min) fournit une forme de définition implicite[2] du prédicat de vérité. Field (2001b), pour sa part, reprend les analyses quiniennes et considère que le prédicat de vérité est transparent. Les énoncés « « A » est vrai » et « A » sont cognitivement équivalents.[3]

Il semble donc que, selon les déflationnistes, la notion de vérité n'ait pas de réel contenu. Si « « A » est vrai » et « A » sont trivialement équivalents, la notion de vérité n'apporte aucun contenu supplémentaire à « A ». Elle ne peut donc pas avoir un rôle explicatif important à jouer dans nos raisonnements.

Les déflationnistes modernes, nous l'avons vu, ne sont pourtant pas éliminativistes, ou partisans de la vérité redondance. Selon eux, le prédicat de vérité joue un rôle expressif crucial dans notre langage, qui le rend à la fois indispensable et inéliminable. Il permet en effet d'asserter indirectement des énoncés auxquels on ne peut avoir un accès direct immédiat. Par exemple, lorsque j'affirme :

(1) La première phrase du *Théétète* est vraie.

Peut-être n'ai-je pas sous la main mon édition préférée des dialogues de Platon et ne suis-je pas en mesure d'aller regarder en quoi consiste précisément cette première phrase. Le prédicat de vérité me permet de remédier à cette incommodité, en m'autorisant une assertion indirecte. Il permet également d'asserter des ensembles infinis d'énoncés, comme l'illustre l'énoncé suivant :

(2) Les théorèmes de l'arithmétique sont vrais.

Cependant, ce caractère indispensable est uniquement dû à l'augmentation de pouvoir expressif que le prédicat de vérité fournit à notre langage, dont les capacités expressives sont limitées du fait de l'absence de quantification substitutionnelle (on ne peut

2. En un sens relâché et non technique, c'est-à-dire non soumis au théorème de Beth. Autrement dit, la théorie minimale, si elle ne fixe pas l'extension du prédicat de vérité dans tout modèle, en fournit néanmoins une axiomatisation satisfaisante.

3. Field, 1994b : la notion de d'équivalence cognitive est ensuite analysée par Field comme un propriété d'interdéducibilité « assez directe » (fairly direct) entre les deux énoncé ; autrement dit du premier énoncé je peux « assez directement » déduire le second et inversement. La notion d'inférence directe devant elle-même être explicitée sans recourir à une notion substantielle de vérité

pas quantifier sur les énoncés, mais uniquement des variables objectuelles [4]) ou d'outils permettant d'exprimer des conjonctions ou des disjonctions infinies [5].

Ainsi, (1) peut s'analyser comme

si la première phrase du *Théétète* est « la neige est blanche » alors la neige est blanche

et

si la première phrase du *Théétète* est « l'herbe est rouge » alors l'herbe est rouge et ...

...

c'est-à-dire comme une sorte de conjonction infinie. Plus précisément, (1) peut s'analyser comme exprimant le même contenu que la conjonction infinie suivante :

$$(1') \bigwedge_{\phi \in \mathcal{L}} (\text{« } \phi \text{ » est la première phrase du Théétète } \rightarrow \phi)$$

De même, (2) peut s'analyser comme

$$(2') \bigwedge_{\phi \in Thm_{PA}} \phi \text{ ou encore } \bigwedge_{\phi \in \mathcal{L}_{PA}} (\ulcorner \phi \urcorner \in Thm_{PA} \rightarrow \phi)$$

Remarquons que dans (1') et dans (2'), le prédicat de vérité n'apparaît plus. Et pour cause, le rôle expressif qu'il jouait n'a plus lieu d'être puisqu'ici on s'est autorisé à utiliser des conjonctions infinies. Pour reprendre une vieille distinction parfois attribuée à Aristote, le prédicat de vérité se rapprocherait donc des expressions syncatégorématiques comme « et », « ou », « existe » qui ne signifient rien en elles-même mais sont indispensables pour combiner des expressions dotées d'un réel contenu (ces dernières étant appelées catégorématiques).

La thèse, d'inspiration déflationniste, plus précise que nous voudrions examiner ici est celle selon laquelle la vérité serait une notion logique ou quasi-logique. Cette thèse n'a que rarement été explicitement ou ouvertement défendue par les principaux auteurs déflationnistes contemporains [6]. Néanmoins, les caractéristiques qu'ils attribuent au pré-

4. Voyez le chapitre précédent pour plus de détails sur cette question.

5. Pour plus de détails sur les liens entre rôle expressif d'un prédicat de vérité déflationniste et conjonctions ou disjonctions infinies, voyez Halbach (1999b).

6. Mais voyez le chapitre précédent et notamment :

> Ce que le minimaliste souhaite néanmoins souligner est que la vérité n'est pas une propriété *complexe* ou *naturelle* [*naturalistic*] mais une propriété d'un autre genre (Field (1992) suggère le terme de « propriété *logique* ») (Horwich, 1998b, p. 37-38, italiques de l'auteur)

Voyez aussi Field (2001b) où Field affirme qu'aux yeux du déflationniste la seule utilité du prédicat de vérité est de remplir

> un important *rôle logique* : il nous permet de formuler certaines conjonctions ou disjonctions infinies que l'on ne pourrait formuler autrement. (Field, 2001b, p. 120, nous soulignons)

dicat de vérité suggèrent fortement un tel rapprochement, notamment lorsqu'ils affirment que le prédicat « vrai » n'est qu'un simple outil logico-syntaxique de décitation permettant l'expression de généralisation, ou lorsqu'ils considèrent que le concept de vérité ne possède aucun pouvoir explicatif propre au sein de nos théories, ou bien encore lorsqu'ils revendiquent que l'analyse correcte de la vérité doit être neutre au regard des questions d'épistémologie ou de métaphysique.

2.2 Logicité

La question est donc posée : la vérité est-elle une notion logique ? Pour tenter de répondre à une telle question, il nous faut être munis d'un critère de logicité ; il faut que nous possédions un moyen de tracer la frontière, ou du moins une frontière, entre les notions logiques et les autres. Cette question a donné lieu à de nombreux travaux et recherches et a été abordée par de nombreux philosophes et logiciens. Malheureusement, le moins que l'on puisse dire, c'est qu'il n'y a pas eu de consensus qui se soit dégagé, tant sur le plan des intuitions fondamentales de ce qui pourrait caractériser les notions logiques, que sur le plan de la méthodologie à employer pour les étudier, ou même sur les résultats techniques précis obtenus, une fois choisie telle ou telle méthodologie.

Il ne peut être question ici de résoudre ces controverses, ni même de faire un panorama complet de toutes les pistes qui ont été explorées autour de ces questions. Notre démarche sera plus modeste, nous allons fixer une méthodologie, introduire un critère de logicité et analyser dans quelle mesure le prédicat de vérité satisfait ou non ce critère. Cela ne constituera donc qu'une réponse partielle à la question posée en titre, mais permettra, nous l'espérons, d'y voir un peu plus clair.

Dans la littérature consacrée à la caractérisation des notions logiques, on peut distinguer deux grandes traditions : d'une part, la tradition inférentielle issue des travaux de Gentzen, et développée, entre autres, par Prawitz et Dummett, d'autre part, la tradition sémantique qui s'appuie sur des outils de théories des modèles. La méthodologie sur laquelle nous allons ici concentrer notre analyse est celle issue des approches preuve-théoriques, ou inférentielles, qui entend étudier les notions logiques à partir du rôle particulier qu'elles jouent dans nos pratiques déductives et dans nos raisonnements. Nous remettons l'étude de l'approche sémantique de la logicité de la vérité à un travail ultérieur.

2.2.1 Caractérisation inférentielle des constantes logiques

En général, les approches inférentielles de caractérisation de la logique s'appuient sur une conception plus large de la signification : les théories de la signification comme usage. C'est en effet un trait essentiel du langage que d'être un « bien public », utilisé par les locuteurs pour communiquer et transmettre de l'information. La signification doit donc pouvoir être partagée et saisie en commun par tous les locuteurs compétents d'une langue. Par conséquent, la signification des énoncés, ou des expressions qui les composent, doit pouvoir être analysée à partir des seuls comportements publiquement observables des locuteurs. Partant de cette intuition centrale, les théories de la signification comme usage considèrent que la signification des mots peut (et doit) être dérivée de, ou même identifiée à, l'usage qu'en font les locuteurs compétents. Un locuteur de français comprend la signification du mot « rouge » dès lors qu'il déploie ou est disposé à déployer certains comportements appropriés : par exemple, placé devant un objet rouge dans des conditions de luminosité normales, un individu maîtrisant le vocabulaire des couleurs en français sera enclin à déclarer « cet objet est rouge ». Si l'on passe sous silence certains raffinements, on peut dire que selon les théories de la signification comme usage, la signification du mot « rouge » en français est identifiée à la classe des comportements linguistiques qui lui sont associés par les membres de la communauté linguistique francophone.

Pour ce qui est des notions logiques, l'usage que nous en faisons consiste principalement à les utiliser dans des inférences. Dans la perspective des théories de la signification comme usage, il a donc été proposé de spécifier les notions logiques comme étant les expressions qui peuvent être caractérisées par un ensemble des règles « purement inférentielles »—ou, comme cela est également parfois dit : « purement structurelles » [7].

Pour illustrer cette idée, prenons l'exemple de la conjonction. Intuitivement, il semble qu'elle puisse être caractérisée par les règles d'introduction et d'élimination suivantes :

Exemple.

$\wedge$-Introduction : $\dfrac{A \quad B}{A \wedge B}$ $\qquad$ $\wedge$-Elimination (G) : $\dfrac{A \wedge B}{A}$ $\qquad$ $\wedge$-Elimination (D) : $\dfrac{A \wedge B}{B}$

Les règles ci-dessus sont purement inférentielles au sens où elles fournissent un schéma d'inférence entièrement caractérisé par sa structure et qui ne dépend pas de la signifi-

7. Nous utiliserons les deux formulations de manière indifférenciée dans ce qui suit.

cation des énoncés qui y apparaissent. En effet, ici les variables A et B sont censées pouvoir être remplacées par n'importe quels énoncés, que ceux-ci portent sur les atomes, les nombres entiers, les éventuels habitants de la planète Mars, ou ce qu'on voudra. Pour maîtriser l'usage de ces règles, il n'est pas nécessaire de comprendre la signification particulière de telle ou telle paire d'énoncés désignés pour prendre la position de A ou de B. Si l'on veut, on peut même imaginer un locuteur francophone monolingue s'amusant à construire des inférences impliquant la conjonction à partir d'énoncés A et B écrits en chinois[8]. Tout individu comprenant la signification du « $\wedge$ », c'est-à-dire de la conjonction, sera capable d'appliquer les règles ci-dessus. Et inversement, un individu ignorant ce qu'est la conjonction pourrait l'assimiler simplement en apprenant à maîtriser ces règles d'inférence, et rien d'autre. Ainsi, l'usage de la conjonction, et partant sa signification, est littéralement constitué par un ensemble de règles d'introduction et d'élimination, purement structurelles.

Par contraste, saisir la signification du mot « rouge », bien qu'elle mette certainement en jeu la capacité à tirer des inférences dans lequel ce terme apparaît, ne peut se limiter à maîtriser un ensemble de règles purement structurelles. De ce contraste est née l'idée que c'était justement le propre des notions logiques que d'avoir une signification entièrement caractérisable au moyen d'un ensemble de règles d'introduction et d'élimination, ou plus généralement au moyen de règles canoniques purement structurelles qui représentent des schémas d'inférences.

Lorsqu'il a introduit le calcul des séquents, Gentzen, 1935 a le premier avancé l'idée que la signification des notions logiques pouvait être donnée par des règles purement structurelles. Commentant de manière informelle, les règles du calcul en déduction naturelle pour la logique classique et la logique intuitionniste, il écrit :

> [...] Les introductions représentent pour ainsi dire les « définitions » des signes qu'elles concernent, et les éliminations ne sont en dernière analyse que des conséquences de ces définition [...] (Gentzen, 1955, p. 27)

Gentzen distingue soigneusement les deux types de règles et c'est aux règles d'introduction qu'il attribue la qualité de 'définitions' de la constante introduite. En effet, pour reprendre notre exemple, lorsque l'on fait des inférences, un usage correct de la conjonc-

8. Sans doute faut-il supposer que ce locuteur est néanmoins capable de reconnaître qu'il s'agit dénoncés, quitte à ce qu'il n'en comprenne pas le sens. Mais une fois cette précision apportée, le point est que ce locuteur peut manipuler la conjonction et effectuer des inférences correctes avec ce connecteur, indépendamment de la signification des conjoints.

tion consistera à n'affirmer $A \wedge B$ que dès lors que l'on a une justification suffisante pour ce faire. Les prémisses de la règle d'introduction exhibent précisément les conditions suffisantes minimales, autrement dit nécessaires, pour pouvoir déduire $A \wedge B$: pour pouvoir conclure $A \wedge B$ il faut et suffit que je sois (au moins) parvenu à A et parvenu à B. De même, pour pouvoir affirmer $A \vee B$ il faut que je soit parvenu à A ou parvenu à B, d'où les règles d'introduction de la disjonction suivantes :

$\vee$-Introduction (G) : $\vee$-Introduction (D) :

$$\frac{A}{A \vee B} \qquad \frac{B}{A \vee B}$$

Ce sont donc bien les règles d'introduction qui confèrent sa signification à la constante concernée.

D'autre part, lorsque Gentzen affirme que les règles d'élimination ne sont en dernière analyse que des conséquences des règles d'introduction, il veut attirer l'attention du lecteur sur le fait que d'après ces règles, on ne pourra inférer d'une prémisse contenant la constante concernée que ce qui était déjà donné comme justification minimale permettant d'introduire cet énoncé : pour pouvoir affirmer $A \wedge B$, je dois avoir (au minimum) justifié A et justifié B. Dès lors, si au cours de mon raisonnement je suis parvenu à $A \wedge B$, je suis justifié à en déduire A et à en déduire B. Dans la suite du passage cité plus haut, Gentzen développe lui-même l'analyse suivante :

> [...] Les introductions représentent pour ainsi dire les « définitions » des signes qu'elles concernent, et les éliminations ne sont en dernière analyse que des conséquences de ces définition, ce que l'on peut exprimer de la façon suivante : dans l'élimination d'un signe, la formule dont il s'agit et dont le signe en question est le signe terminal ne peut « être utilisée que dans le sens que lui confère l'introduction de ce signe ». (Gentzen, 1955, p. 27)

Cette relation entre règles d'introduction qui fixent la signification du signe concerné et les règles d'élimination qui ne permettent de déduire d'un énoncé contenant le signe concerné que ce qui a déjà dû être minimalement établi lorsque celui-ci a été introduit, est connu depuis Prawitz (1965) sous le nom de Principe d'Inversion.

> [...] une règle d'élimination est, en un sens, l'inverse de la règle d'introduction correspondante : par l'application d'une règle d'élimination, on ne fait essentiellement que restaurer ce qui avait déjà été établi si la prémisse majeure de l'application avait été établie par l'application d'une règle d'introduction. (Prawitz, 1965, p. 33)

Ainsi Gentzen remarque que les notions logiques qu'il introduit dans son système de déduction naturelle sont caractérisables (ou, nous dit-il, « définissables ») au moyen de règles structurelles. L'objectif à proprement parler de son travail était de développer des systèmes de preuves pour la logique du premier ordre. Il n'était pas à la recherche d'un critère de logicité. Pour autant, les idées contenues dans sa remarque ont eu une très riche postérité philosophique. De nombreux auteurs s'en sont emparés et l'ont dotée d'une dimension normative. Les constantes logiques considérées par Gentzen ont une signification donnée par des règles purement structurelles, mais ce n'est pas le fruit du hasard. C'est au contraire une propriété spécifique : seront logiques uniquement les notions qui peuvent être ainsi caractérisées.

Toutefois, l'idée selon laquelle toute notion caractérisable au moyen d'un ensemble de règles d'introduction et d'élimination serait une notion logique s'est très vite heurtée à un problème. Contre cette idée, Prior (1960) a proposé un fameux contre- exemple en introduisant les règles suivantes permettant de définir une constante « pathologique », la constante « *tonk* » :

tonk-Introduction :	*tonk*-Elimination :
$\dfrac{A}{A \ tonk \ B}$ *tonk*-Intro	$\dfrac{A \ tonk \ B}{B}$ *tonk*-Elim

Ici, la constante putative « tonk » est bien définie par des règles purement structurelles. Néanmoins, on voit aisément que si on ajoute ce connecteur « tonk » à un système de preuve cohérent, on court à la catastrophe. Si le système d'origine permet de dériver au moins un énoncé de manière inconditionnelle (ce peut être un axiome logique ou un énoncé dérivable sans hypothèse par simple application des règles d'inférence du système) son extension par « tonk » donne aussitôt une système incohérent, permettant de dériver n'importe quel énoncé.

Soit B l'énoncé « dieu existe ». Si notre système d'origine contient la règle standard pour l'introduction de $\rightarrow$, on prouve alors l'existence de dieu comme suit :

$$\dfrac{\dfrac{\dfrac{[A]}{A}}{(A \to A)}\,{\to}\text{-Intro}}{(A \to A)\ tonk\ B}\,tonk\text{-Intro}}{B}\,tonk\text{-Elim}$$

Cependant, ce contre-exemple ne signifie pas qu'il faille nécessairement abandonner l'idée de donner un critère de logicité fondé sur le rôle particulier que jouent les notions logiques dans nos inférences. Pour une notion donnée, le fait d'être caractérisable par des règles structurelles n'est peut-être qu'une condition nécessaire mais non suffisante pour pouvoir être qualifiée de logique. Peut-être faut-il y adjoindre des contraintes supplémentaires, dûment motivées, qui permettraient d'écarter les cas pathologiques. Plusieurs propositions ont été faites en ce sens.

Dans *The Logical Basis of Metaphysics* (1991), Dummett s'attaque frontalement au problème : il affirme que pour qu'une notion soit logique, il ne suffit pas de pouvoir en donner une caractérisation par un ensemble de règles d'introduction et d'élimination arbitraires. Encore faut-il que les règles d'introduction et d'élimination soient, selon l'expression qu'il introduit, en *harmonie*. Partant de cette notion informelle, Dummett avance deux conditions supplémentaires qu'une constante logique doit satisfaire. Il distingue en fait deux critères : l'harmonie globale (condition de conservativité) et l'harmonie locale (condition dite de réduction ou de normalisation). À la suite de Belnap (1962), il propose le critère suivant :

Critère d'harmonie globale. (Conservativité) L'ajout d'une nouvelle constante logique au moyen de règles purement structurelles doit produire une extension conservative.

Précisons ce que cela signifie. Supposons que nous partions d'un système logique **S** couché dans un langage $\mathcal{L}$. Notre système **S** est muni de règles d'inférences gouvernant l'usage des constantes logiques présentes dans $\mathcal{L}$. Supposons en outre que l'on étende $\mathcal{L}$ en un langage $\mathcal{L}'$ contenant un nouveau symbole δ, dont la signification est donnée par de nouvelles règles d'introduction et d'élimination. On obtient alors un système étendu **S'** qui contient les règles données par **S** ainsi que les nouvelles règles gouvernant δ. **S'** est une extension conservative de **S** si tout énoncé de ϕ de $\mathcal{L}$ (qui donc ne contient aucune occurrence du nouveau symbole δ) que l'on peut dériver dans **S'** (au moyen d'une déduction qui elle contiendra peut-être des énoncés, autres que ϕ , dans lesquels

δ apparaît) est déjà dérivable dans **S**.[9] Autrement dit, tout énoncé du langage d'origine prouvable dans le langage étendu devait déjà l'être dans le langage d'origine.

A ce critère, s'en ajoute un second :

Critère d'harmonie locale. (Réduction) Les règles d'introduction et d'élimination fixant la signification d'une constante logique δ doivent être telles que toute formule maximale contenant δ en position principale doit pouvoir être réduite.

Une formule maximale est une formule qui apparaît dans une déduction comme conséquence d'une règle d'introduction et comme prémisse majeure d'une règle d'élimination.
À titre d'illustration voici comment fonctionne la réduction sur une formule maximale dont le connecteur principal est une conjonction.

$$\dfrac{\dfrac{\begin{matrix}\Pi_1 \\ A\end{matrix} \quad \begin{matrix}\Pi_2 \\ B\end{matrix}}{A \wedge B}\,\wedge\text{-Intro}}{A}\,\wedge\text{-Elim} \quad \Rightarrow \quad \dfrac{\Pi_1}{A}$$

Il est clair que les règles pour *tonk* ne satisfont pas ces contraintes d'harmonie supplémentaires. Comme le montre la « preuve » de l'existence de Dieu, elles ne sont ni réductibles ni conservatives. À la suite des travaux de Dummett et Prawitz, nous pouvons donc proposer le critère inférentialiste de logicité suivant :

Critère inférentialiste de logicité. Une expression est logique si sa signification est déterminée par des règles purement structurelles et qui vérifient les critères de conservativité (harmonie globale) et de réduction (harmonie locale).

2.3 Une caractérisation inférentielle de la vérité ?

2.3.1 Première approche

Ce critère va nous permettre d'aborder la question de la logicité de la vérité de manière plus précise. La thèse quelque peu vague selon laquelle le prédicat de vérité est

9. En toute rigueur, il convient de distinguer plusieurs types de conservativité : la conservativité syntaxique ou déductive, et la conservativité sémantique. Bien entendu, en bon intuitionniste, c'est la notion de conservativité déductive que Dummett a en vue. De toute façon, pour la logique du premier ordre classique ou intuitionniste, les deux notions coïncident. Les choses sont différentes pour les logiques d'ordre supérieur, ou plus généralement pour les logiques qui ne sont pas munies d'un système de preuve complet.

une « sorte » de constante logique trouve ici un cadre dans lequel elle peut être examinée de manière rigoureuse. Il nous faut voir dans quelle mesure le prédicat de vérité satisfait notre critère inférentialiste de logicité. Mais d'emblée un problème se pose. Pour pouvoir tester la logicité du prédicat de vérité en utilisant la méthodologie que nous venons de décrire, il faut proposer des règles putatives censées gouverner le prédicat de vérité puis vérifier si elles satisfont le critère de logicité avancé. Or, faire cela c'est déjà prendre un parti : dans le cas des constantes logiques du premier ordre, il semble sans doute plausible à première vue que les règles proposées fixent entièrement la signification des constantes concernées. Mais pour le prédicat de vérité l'existence de telles règles est bien plus sujette à caution. Les partisans d'une théorie récursive axiomatisée à la Tarski ou d'une théorie substantielle de la vérité (vérité cohérence ou vérité correspondance par exemple) douteront d'emblée que de telles règles puissent exister. Avec cet avertissement en tête, nous pouvons néanmoins proposer des éléments de réflexion.

Les auteurs déflationnistes ont la plupart du temps [10] présenté leur théorie de la vérité non pas sous la forme d'une paire de règles d'introduction et d'élimination mais sous la forme d'une théorie axiomatique, composée de la collection des instances du schéma-T de Tarski :

$$a \text{ est vrai} \leftrightarrow \text{A}$$

où A est un énoncé du langage objet, c'est-à-dire un énoncé dans lequel le prédicat de vérité n'apparaît pas et où a est un nom de l'énoncé A obtenu soit par une mise entre guillemets soit par une description structurelle faisant intervenir une théorie de la syntaxe du langage objet . Cette collection de biconditionnels est censée gouverner l'emploi du prédicat « vrai », et dire « tout ce qu'il y a à dire » sur la vérité. On retrouve ici l'idée que la signification de cette notion peut être saisie à partir de ressources très modestes. Dans le cas présent, celles-ci sont présentées sous la forme de biconditionnels ; néanmoins, le passage à une version inférentielle de la théorie déflationniste ne pose pas véritablement de problème. Les biconditionnels permettent de déduire directement l'énoncé « a est vrai » de l'énoncé « A » et inversement. Ceci suggère donc tout naturellement la paire

10. Hartry Field constitue une exception notable sur ce point, puisque dans certains textes il formalise la théorie déflationniste de la vérité en posant que pour tout énoncé A du langage objet, A et $Vr(\text{« } A \text{ »})$ sont « cognitivement équivalents »—l'équivalence cognitive de deux expressions pour un agent donné, étant définie par lui comme la possibilité de déduire de manière relativement directe (fairly direct) une expression à partir de l'autre et inversement. Ainsi l'équivalence entre « A » et « Vr(« A ») » est bien donnée sous la forme de *règles* d'inférence.

de règles suivantes, censées gouverner entièrement l'emploi de « vrai » et déterminer sa signification :

Définition. *Règles Générales pour « Vrai »* :

$Vrai$-Introduction : $\dfrac{A}{Vr(a)}\ Vr\text{-Intro}$

$Vrai$-Elimination : $\dfrac{Vr(a)}{A}\ Vr\text{-Elim}$

où « A » doit être remplacé par un énoncé du langage considéré, et « a » doit être remplacé par un nom ou un terme désignant cet énoncé.

Si, d'une part, on se cantonne à un prédicat de vérité typé [11] et si, d'autre part, le système logique auquel on ajoute des règles pour la vérité contient une flèche d'implication gouvernée par les règles habituelles [12] alors les deux formulations sont équivalentes : à partir de la collection des biconditionnels $Vr(a) \leftrightarrow A$, on montre par simple application du modus ponens que les règles Vr-Intro et Vr-Elim sont admissibles, et inversement une fois munis de ces deux règles on peut facilement déduire chacun des biconditionnels dont la collection forme la version axiomatisée de la théorie déflationniste [13] Examinons le cas d'un prédicat de vérité *typé* dont l'emploi est gouverné par les règles d'inférences ci-dessus. Les règles proposées pour la vérité sont-elles logiques ? Autrement dit, satisfont-elles notre critère de logicité ? Sont-elles purement structurelles, et *harmonieuses* ? De prime abord, il semble que non.

11. C'est-à-dire ne portant que sur des (noms d') énoncés qui ne contiennent pas le prédicat « Vrai »
12. À savoir, la règle d'introduction

$$\rightarrow\text{-Intro}\ \frac{\begin{array}{c}[A]\\ \vdots \\ B\end{array}}{A \rightarrow B}$$

et la règle d'élimination par modus ponens

$$\frac{A \qquad A \rightarrow B}{B}\ \rightarrow\text{-Elim}$$

13. Notons que les deux formulations de la théorie déflationniste (par un couple de règles ou par une axiomatisation au moyen d'une collection infinie de biconditionnels) ne sont pas équivalentes lorsqu'on passe au cas non typé (où le prédicat de vérité peut s'appliquer à des noms d'énoncés qui eux-mêmes contiennent déjà une occurrence du prédicat de vérité). Mais nous n'entreprendrons pas l'étude de cette question ici.

Certes, les règles proposées satisfont le critère d'harmonie locale. Pour toute formule maximale contenant le prédicat « Vr » en position principale, la réduction s'opère de la façon suivante :

$$\dfrac{\dfrac{\begin{matrix}\Gamma\\ \Pi_1\\ A\end{matrix}}{Vr(a)}\,\text{Vr-Intro}}{A}\,\text{Vr-Elim} \quad \Rightarrow \quad \dfrac{\begin{matrix}\Gamma\\ \Pi_1\\ A\end{matrix}}{A}$$

En revanche, les autres conditions de notre critère de logicité, à savoir la pure structuralité des règles et l'harmonie globale, posent problème. En effet, l'utilisation des règles proposées nécessite d'introduire des noms d'énoncés sur lesquels portera le prédicat de vérité. Tacitement, adopter les règles pour la vérité, c'est donc accepter un certain engagement ontologique quant à l'existence de toute une classe d'objets, à savoir la collection des noms d'énoncés sur lesquels le prédicat de vérité pourra porter. Or, traditionnellement, on considère que la logique doit rester neutre quant aux questions d'existence. [14] Si accepter les règles pour la vérité, c'est accepter l'existence d'une infinité d'objets nouveaux, les noms d'énoncés du langage objet, ou accepter l'existence d'une théorie de la syntaxe permettant de construire ces objets, il semble que cela nécessite des ressources qui dépassent celles de la pure logique. Comme on peut s'y attendre, cet engagement ontologique implicite se traduit par des résultats de non conservativité, et le critère d'harmonie globale n'est pas satisfait par la notion déflationniste de vérité, telle qu'elle est ici caractérisée. Plus spécifiquement, Halbach, 2001b a le premier souligné que les théories déflationnistes de la vérité, si modestes soient-elles, n'étaient pas conservatives sur la logique du premier ordre pure avec égalité. La preuve donnée par Halbach concerne la version axiomatisée par les T-équivalences des théories déflationnistes, mais elle est aisément adaptable à notre cadre. [15] Elle repose sur le fait que les règles pour la vérité et le principe de substitution des identiques « forcent » que tout énoncé prouvable inconditionnellement (c'est-à-dire toute validité) et sa négation doivent recevoir des noms distincts. Dans le système étendu par Vr-Intro/Elim, on peut alors prouver l'énoncé

14. Encore que, en théorie des modèles, l'interprétation standard de la logique du premier ordre ne prend en considération comme structure d'interprétation que les structures dont le domaine d'objets n'est pas vide, ce qui fait de l'énoncé '$\exists x(x = x)$' une vérité logique. De ce point de vue, la bonne logique, la logique la plus neutre au regard des questions ontologiques, devrait être la logique libre, qui admet les domaines d'interprétation vides et dans laquelle l'énoncé qui précède n'est pas toujours valide.

15. Voir l'appendice pour une démonstration en bonne et due forme.

$\exists x \exists y (x \neq y)$, qui affirme l'existence d'au moins deux objets distincts. Cet énoncé est un énoncé du langage de la logique du premier ordre avec égalité, mais ce n'est pas un théorème de la logique du premier ordre pure avec égalité : il n'est pas prouvable sans recourir aux règles pour la vérité ; et les règles pour « vrai » ne sont donc pas conservatives sur la logique.

Ce processus de nominalisation qui préside à l'introduction du prédicat de vérité pose également problème au regard du caractère purement structurel des règles proposées. Hodes (2004) l'un des rares auteurs à avoir proposé une analyse de la notion de vérité à partir d'une approche preuve-théorique de la logicité souligne que l'usage du prédicat de vérité s'appuie en fait sur une information sémantique. En effet, les inférences du type

$$\frac{\text{la neige est blanche}}{a \text{ est vrai}} \quad \text{et} \quad \frac{a \text{ est vrai}}{\text{la neige est blanche}}$$

nécessitent de connaitre la référence du nom « a » comme étant l'énoncé « la neige est blanche ». Dans certains cas cette connaissance peut paraître triviale ou immédiate, comme lorsque lorsqu'on a sous les yeux le nom d'un énoncé obtenu par mise entre guillemets. Dans d'autres cas, l'application de ces règles peut mettre en jeu des ressources considérables : que l'on songe, par exemple, à un prédicat de vérité formalisé qui s'applique à des codes d'énoncés à la Gödel. Pour retrouver, l'énoncé derrière tel ou tel entier je dois recourir à la machinerie récursive permettant le décodage. Dans d'autres cas encore, elle peut requérir une vaste faisceau de connaissances de nature diverse. Si je sais que β est un nom de la dernière phrase prononcée par Napoléon lors de son allocution aux troupes, le 21 Juillet 1798 à Embabèh, de « Vr(β) », que puis-je déduire exactement ? Pour pouvoir appliquer ici la règle d'élimination et inférer l'énoncé « Du haut de ces pyramides quarante siècles vous contemplent », je dois pouvoir déterminer la référence de β et, en l'espèce, bien connaître mon histoire de France et les citations de ses hommes illustres.

On pourrait multiplier à l'envi les exemples où connaître la référence du nom de l'énoncé auquel s'applique le prédicat « vrai », est particulièrement difficile et réclame des connaissances de nature mathématique, empirique ou historique, *etc.* Cette information indispensable à l'application des règles d'introduction ou d'élimination est de nature sémantique. Le point crucial, selon l'analyse de Hodes, est que les règles pour la vérité reposent sur un processus de nominalisation des énoncés, qui contient une connaissance sémantique [16] (que cette dernière soit triviale ou complexe n'y change rien). Elles ne sont

16. à savoir, la connaissance de la référence du nom auquel le prédicat « vrai » s'applique.

donc *pas* purement structurelles. Et, la vérité, même si l'on admet que sa signification est fixée par les règles proposées, ne peut donc pas, *stricto sensu*, être qualifiée de logique [17].

En résumé, les règles déflationnistes censées fixer la signification de « vrai » ne paraissent pas, à première vue, satisfaire le critère de logicité avancé dans les caractérisations preuve-théoriques des notions logiques. Certes, elles vérifient bien le critère d'harmonie locale, puisqu'il existe en effet une procédure permettant de réduire toutes les formules maximales où le prédicat de vérité apparaît en position principale. Mais, les règles proposées exigent de faire un détour par des noms d'énoncés, ce qui induit un engagement ontologique vis-à-vis de ces entités et requiert d'être en possession d'une information d'ordre sémantique. Elles ne sont donc ni purement structurelles ni conservatives sur la logique du premier ordre.

2.3.2 Raffinements

Pour autant, la question n'est peut-être pas définitivement tranchée. Dans un travail récent Henri Galinon [18] a proposé une analyse inférentielle du prédicat de vérité un peu différente de celle de Hodes, et défend la thèse déflationniste selon laquelle la vérité est une notion logique ou quasi logique. Puisque les règles proposées pour la vérité satisfont le critère de réduction, l'analyse va devoir se concentrer sur les deux autres conditions qui composent le critère de logicité : l'harmonie globale, comprise comme clause de conservativité, et le caractère purement structurel des règles.

2.3.2.1 La conservativité comme critère de logicité

Réexaminons tout d'abord, le phénomène de la non-conservativité de la vérité déflationniste sur la logique pure (avec égalité). Pour proposer une défense de la thèse selon laquelle la vérité est une notion logique, pour tenter de « sauver » la logicité de la vérité malgré son « échec » de conservativité, on peut suivre un double mouvement : tout d'abord, on peut chercher à remettre en cause le critère de conservativité lui-même, et ce faisant minimiser la portée du résultat négatif qui frappe les règles déflationnistes. Parallèlement, on peut également essayer de montrer que les règles pour « Vrai » sont « presque » conservatives, qu'il s'en faut de très peu qu'elles ne remplissent la condi-

17. Signalons tout de même que dans son article, Hodes suggère de qualifier la vérité de notion « quasi-logique ».

18. Henri Galinon, *Recherches sur la Vérité—Définition, Élimination, Déflation*, Thèse de doctorat, 2010.

tion d'harmonie globale, peut-être sous réserve d'hypothèses auxiliaires. C'est, ce nous semble, ce type de stratégie argumentative qui guide le travail d'Henri Galinon sur cette question. [19] Nous en reprenons ici les grandes lignes.

Nous ne tenterons pas de proposer un traitement exhaustif de la question du rôle que peut ou doit jouer la conservativité comme critère de logicité. Ce problème fait partie des discussions, toujours en cours, opposant les divers partisans d'une caractérisation inférentielle des constantes logiques [20]. Néanmoins, il nous paraît indispensable de donner quelques indications à ce sujet, au regard de la question de la logicité de la vérité.

Comme nous l'avons signalé dès le début de ce chapitre, il n'existe pas, même au sein des partisans d'une approche inférentialiste de la logicité, de consensus général sur le bon critère de logicité. Il semble néanmoins que la plupart des auteurs s'accordent sur le fait que la signification d'une expression logique doit pouvoir être donnée par des règles purement structurelles qui soient en harmonie locale. À l'inverse, la clause de conservativité, après avoir été introduite pour la première fois par Belnap en réponse au problème posé par *tonk*, puis ardemment défendue par Dummett, a soulevé bien des objections. On peut distinguer trois grands types de critiques adressées à ce que Dummett a baptisé l'harmonie globale. Une première difficulté soulevée par cette notion tient au fait que la conservativité n'est pas tant une propriété propre à un couple de règles pris isolément, mais plutôt une propriété plus large reliant un système logique **S** et son extension **S** $\subseteq$ **S'**. Un même ensemble de règles, censées réguler l'usage d'une expression δ, pourra donc donner une extension conservative lorsqu'on l'ajoute à un premier système logique $\mathbf{S}_1$, mais donner une extension non-conservative lorsqu'on l'ajoute à un autre système $\mathbf{S}_2$ [21]. C'est ce qu'on pourrait appeler le problème de l' « instabilité » de l'harmonie globale. Une seconde difficulté découle du fait que le caractère conservatif ou non d'une extension par un ensemble de règles pour une expression δ est également sensible au cadre dans lequel on formalise nos inférences. Nous avons choisi ici un cadre de déduction naturelle ; mais il existe d'autres possibilités —en particulier le calcul des séquent— et les résultats de conservativité ne seront pas forcément les mêmes lorsqu'on passe d'un cadre à un autre. Enfin, en troisième lieu, on peut remarquer qu'une application trop stricte du

19. *cf.* Galinon (2010, p. 309-335).

20. Pour une défense et élaboration de l'harmonie globale, voir Dummett, 1973 et surtout Dummett, 1991. Pour une critique de la notion Dummettienne de conservativité, voir Read, 2000, 1988, ainsi que Milne, 1994.

21. Si le système étendu $\mathbf{S}_1 \cup \{\text{règles pour } \delta\}$ est conservatif sur $\mathbf{S}_1$, alors que le système étendu $\mathbf{S}_2 \cup \{\text{règles pour } \delta\}$ n' est pas conservatif sur $\mathbf{S}_2$, que doit-on dire des règles pour δ ? Sont-elles oui ou non en harmonie globale ?

critère de conservativité conduit à exclure certaines notions qui, pourtant, paraissent indéniablement logiques.

Commençons par le problème de l'instabilité de l'harmonie globale. Lorsqu'on ajoute à un système logique un nouveau symbole muni de nouvelles règles, le caractère conservatif ou non de l'extension obtenue va bien souvent dépendre du système d'origine. Il semble donc que la conservativité ne soit pas une propriété intrinsèque des règles d'introduction et d'élimination, mais plutôt une propriété globale reliant deux systèmes logiques dont l'un est une extension de l'autre. Ainsi, les règles pour *tonk*, nous l'avons déjà noté, engendrent un système incohérent où tout énoncé devient aussitôt dérivable. Les règles sont donc non-conservatives sur tout système « raisonnable » auquel on voudrait les ajouter. Mais, remarquons que si le système d'origine, dénué de toute *tonkerie*, est déjà lui-même incohérent, alors l'extension par *tonk* sera trivialement conservative.

Au delà de ce cas pathologique, le phénomène d'instabilité de la conservativité pose de sérieux problèmes comme le montre cet autre exemple, plus intéressant. Considérons un système de déduction naturelle pour la Logique Positive **LP**, c'est-à-dire un système contenant les règles pour « $\wedge$ », « $\vee$ », « $\rightarrow$ », « $\forall$ », et « $\exists$ » [22] mais ne contenant pas de négation. Supposons qu'on étende ce système en lui ajoutant des règles classiques pour la négation « $\neg$ ». On obtient alors une extension non conservative [23]. Les lois classiques suivantes, dans lesquelles aucune occurrence de « $\neg$ » n'apparaît, ne sont en effet pas prouvables à partir des seules règles de **LP** :

$$\vdash_{LC} ((p \rightarrow q) \rightarrow p) \rightarrow p \qquad \text{(Loi de Peirce)}$$
$$\vdash_{LC} \exists x(\exists y Fy \rightarrow Fx) \qquad \text{(Seconde Loi de Peirce)}$$

Cependant, ces lois deviennent prouvables dès lors qu'on étend **LP** par les règles classiques pour la négation. Ainsi, ces règles données pour « $\neg$ » et censées en fixer la signification ne sont pas conservatives sur le fragment positif de la logique classique [24]. Par contraste, si au lieu de partir de **LP**, le fragment positif de la logique, on prend comme point de départ un système ne contenant que la conjonction auquel on ajoute la négation classique, on obtient une extension conservative. Il n'existe pas d'énoncé contenant comme seule constante logique « $\wedge$ » qui serait dérivable à partir des règles pour « $\wedge$ »

22. Voir l'appendice technique pour une formulation explicite.
23. **LC**, la logique classique, n'est pas conservative sur **LP**
24. Il s'agit là d'un résultat bien connu. Nous empruntons cet exemple appliqué à une discussion du critère de conservativité à Read (2000, 1988), repris dans Galinon (2010).

et pour « ¬ », sans l'être déjà à partir des seules règles pour « ∧ ». En résumé, dans un système de déduction naturelle semblable à ceux que nous considérons ici, les règles pour « ¬ » donneront ou non une extension conservative selon le « moment » où on les ajoute aux règles gouvernant les autre constantes, ou, pour le dire autrement, selon l' « ordre » dans lequel les expressions et les règles qui les gouvernent sont introduites. Si l'on part d'un système ne contenant que la conjonction et qu'on l'étend par une négation classique, on obtient une extension conservative. Si l'on construit tout d'abord le fragment positif de la logique et qu'ensuite on lui ajoute une négation classique, cette fois, l'ajout de la négation produit une extension non-conservative.

Est-ce là une raison suffisante pour considérer que la négation, telle qu'elle est interprétée par les logiciens classiques, n'est pas une notion logique ? Sur ce point les avis divergent. Dummett et Prawitz en tirent un argument contre la logique classique et affirment que la seule véritable logique, celle qui rend justice à l'idée que la signification des constantes est données par des règles purement structurelles en harmonie, est la logique intuitionniste. Mais on peut également voir dans le phénomène précédent une forme de *reductio* du critère de conservativité. Pour qui considère comme incontestable que la négation classique est une notion logique, le fait que ses règles d'introduction et d'élimination produisent une extension non-conservative de **LP** montre simplement que la condition d'harmonie globale, telle quelle est défendue par Dummett, n'est pas adéquate.

À vrai dire, l'origine de ce phénomène d'instabilité réside dans le fait que les règles de la logique classique en déduction naturelle ne sont pas *séparables*. Les règles d'un système logique sont dites séparables si tout énoncé valide du système est dérivable à partir des seules lois gouvernant les constantes logiques qui ont une occurrence dans cet énoncé, sans qu'il soit nécessaire de faire appel aux règles gouvernant les autres constantes. La propriété de séparabilité est liée à celle de conservativité, mais elle est un peu plus forte. Si un système logique jouit de la propriété de séparabilité, alors pour chaque constante logique du système, les règles fixant son usage sont conservatives sur *tout* sous-fragment du système considéré. Par conséquent, dans une telle situation, chaque ensemble de règles gouvernant une constante satisfera le critère de conservativité, quel que soit l' « ordre » dans lequel on choisit d'introduire les constantes. En déduction naturelle, les règles de la logique minimale **LM**, tout comme celles de la logique intuitionniste **LI**, sont séparables. En revanche, lorsqu'on passe à la logique classique **LC**, par exemple en ajoutant la règle

d'élimination des doubles négations, les règles ne sont plus séparables, comme l'illustre l'exemple des lois de Peirce.

Le problème est d'autant plus épineux que la non-conservativité de la négation classique sur la logique positive dépend également du cadre dans lequel sont formalisées les inférences. Nous avons privilégié ici un cadre de déduction naturelle. Mais si l'on avait choisi plutôt de formaliser nos inférences dans un système de calcul des séquents à la Gentzen, on obtiendrait des résultats différents. En particulier, dans un format de calcul des séquents multiconclusion muni des règles standard pour les constantes de la logique du premier ordre, les règles pour la négation classique sont conservatives sur le fragment positif de la logique. Et, certains systèmes de calcul des séquents pour la logique classique vérifient la propriété de séparabilité. [25] Cette sensibilité des propriétés de conservativité au cadre de formalisation des inférences constitue une autre limite à la portée normative que cette condition peut avoir dans les tentatives de caractérisation de la logicité. Bien souvent, les philosophes et logiciens dont le coeur penche en faveur de la logique classique privilégient un cadre d'analyse des inférences fondé sur le calcul des séquents (multiconclusion) tandis que les adeptes de la logique intuitionniste lui préfèrent la déduction naturelle [26].

Quoi qu'il en soit, et quoi que l'on pense de la querelle opposant les partisans d'une logique intuitionniste aux défenseurs de la logique classique, la négation n'est pas le seul cas problématique. Même pour des notions dont la logicité ne semble pas prêter à controverse, on peut construire des systèmes logiques « exotiques » où des phénomènes de non conservativité apparaissent. Dummett lui-même propose un exemple de ce type portant sur la disjonction. Il considère un système **S** ne contenant que la conjonction standard « $\wedge$ » et une notion affaiblie de disjonction, notée « $\dot{\cup}$ », pour laquelle la règle d'élimination ne peut pas s'employer sous hypothèses non déchargées. Lorsqu'on étend **S** en lui ajoutant la disjonction standard « $\vee$ », gouvernée par les règles habituelles, on obtient un système etendu **S'** = **S** $\cup\{\vee\}$ dans lequel on peut dériver une loi de distributivité pour « $\wedge$ » et « $\dot{\cup}$ », à savoir : $p \wedge (q \mathbin{\dot{\cup}} r) \vdash_{S'} (p \wedge q) \mathbin{\dot{\cup}} (p \wedge r)$. Or, cette loi n'est pas prouvable dans le système d'origine **S** [27]. Ainsi, les règles habituelles (non affaiblies) pour la disjonction « $\vee$ » sont non conservatives sur le système **S**. Néanmoins,

25. *Cf.* par exemple, Troelstra et Schwichtenberg (2001, p. 55-57).

26. Bien entendu, la question du choix du bon cadre d'analyse de nos pratiques inférentielles n'est pas qu'affaire de goût. Elle est abondamment discutée par les tenants d'une approche inférentielle ou preuve-théorique de la logicité. Nous ne pouvons malheureusement pas en discuter plus avant ici.

27. *Cf.* Dummett (1991, p. 287-290), pour plus de détails.

il semble difficilement acceptable de devoir bannir la disjonction standard du royaume des notions logiques.

Bref, on le voit, la condition de conservativité, l'harmonie globale selon la terminologie de Dummett, est loin de constituer un critère de logicité non problématique et universellement accepté. Notons néanmoins qu'un abandon pur et simple de cette contrainte ne constituerait pas forcément une solution satisfaisante. Si l'on se contentait d'expurger purement et simplement notre critère de logicité de la condition de conservativité, c'est-à-dire si l'on qualifiait de logique toute expression dont la signification est caractérisée par des règles purement structurelles et obéissant à une procédure de réduction (i.e. en harmonie *locale*), on s'exposerait à la désagréable conséquence de considérer comme logiques des expressions incohérentes produisant, à l'instar de *tonk*, un système contradictoire. Ce danger a été signalé par Read (2000) [28]. Nous touchons là un point limite des tentatives de caractérisation de la logique, inspirées par les approches inférentialistes de la signification.

Dès lors, concernant la vérité, peut-être le caractère conservatif de l'extension n'est-il pas si fondamental ? Peut-être ne devrait- on pas dénier à la vérité une nature logique au seul motif qu'elle n'est pas conservative sur la logique du premier ordre et ne satisfait donc pas l'harmonie globale ? D'ailleurs, et c'est là le deuxième versant d'une défense de la logicité de la vérité, on peut remarquer que si l'extension par les règles d'introduction et d'élimination du prédicat de vérité n'est certes pas conservative sur la logique du premier ordre avec égalité, cet échec de conservativité est pour ainsi dire « minime ».

Nous avons effectivement vu que les règles pour « Vrai » nécessitaient d'introduire des noms pour chaque énoncé du langage et que cela s'accompagnait d'un engagement ontologique vis-à-vis de ces entités. De cet engagement implicite découlent de nouveaux énoncés existentiels qui ne sont pas dérivables à partir des seules ressources de la logique pure, ce qui entraîne la non-conservativité de la vérité. Mais, pourrait-on dire, ce qui est à l'origine de cette non-conservativité, c'est l'acceptation de l'existence de noms (distincts) pour les différents énoncés du langage considéré. C'est parce que je veux pouvoir nommer (et distinguer), par exemple, l'énoncé « A » et l'énoncé « ¬A », que je dois introduire deux objets nouveaux : un nom différent pour chacun de mes énoncés —ou bien encore une théorie de la syntaxe me permettant de les construire. La vérité n'a rien à voir là dedans. Il ne faut pas lui en tenir rigueur, ni lui attribuer une quelconque

28. *Cf.* ce que Read appelle « Le menteur preuve-théorique » (« *A Proof-conditional Liar* »), Read (2000, p. 141).

substantialité pour autant. D'ailleurs, je puis tout à fait vouloir étudier les énoncés d'un langage $\mathcal{L}$ et devoir pour ce faire les nommer ou disposer d'une théorie de la syntaxe de $\mathcal{L}$, indépendamment de toute considération sémantique ou de toute notion de vérité. Que l'on songe, par exemple, à la caractérisation inductive des expressions bien formées (i.e. la spécification des suites ou séquences de symboles d'un alphabet qui sont des formules) donnée dans les manuels de logique lorsqu'on introduit un langage formel : ici, il n'est pas question de valeur de vérité, ou d'interprétation des formules, et c'est à juste titre que ce type de chapitre est appelé syntaxe (par opposition à la sémantique, dont dépendent les questions traitant de la vérité).

C'est pourquoi, dans l'analyse du phénomène de non-conservativité de l'extension de la logique par des règles pour la vérité, il faut séparer ce qui relève d'une théorie de la syntaxe et ce qui relève de la vérité proprement dite. Or, justement, les règles déflationnistes pour la vérité sont conservatives sur une théorie de la syntaxe : [29]

Théorème. Soit T une théorie contenant sa propre syntaxe, par exemple une théorie contenant un peu d'arithmétique, telle que Q ou PA. Alors,

$$T \cup \{Vr(\ulcorner \varphi \urcorner) \leftrightarrow \varphi : \varphi \in \mathcal{L}_T\}$$

est conservative sur T.

Une fois acceptée l'existence des noms d'énoncés, auxquels le prédicat de « vrai » peut être accolé, nous avons déjà dépassé la simple logique, nous sommes déjà placés dans une théorie non-conservative sur la logique pure ; mais l'ajout subséquent d'un prédicat de vérité gouverné par les règles déflationnistes n'aura en revanche pas d'impact ontologique supplémentaire. Quel enseignement peut-on tirer de ce résultat concernant la nature plus ou moins logique de la vérité ?

Les déflationnistes auraient-ils parlé trop vite en affirmant que la vérité était « une [sorte de] notion logique » ? Pas forcément. Il est vrai que les règles pour « Vrai » ne satisfont pas *stricto sensu* l'harmonie globale. Mais le théorème précédent suggère que lorsqu'on étend un système logique au moyen des règles générales :

29. Ce résultat bien connu est dû à l'origine à Tarski —voyez le chapitre précédent section 1.1.4—, sous la forme d'un résultat de conservativité de l'ensemble des T-équivalences sur l'arithmétique. On trouvera une démonstration de ce résultat adaptée de Halbach, 2014 dans l'appendice.

$Vrai$-Introduction :	$Vrai$-Elimination :	où « A » doit être remplacé par un énoncé du langage considéré, et « a » doit être remplacé par un nom de cet énoncé.
$\dfrac{A}{Vr(a)}\ Vr\text{-Intro}$	$\dfrac{Vr(a)}{A}\ Vr\text{-Elim}$	

on peut distinguer deux processus. Une première étape consiste à se donner les ressources syntaxiques nécessaires pour pouvoir nommer les expressions du langage de notre système. Pour chaque A, il nous faut disposer d'un nom a. Faire cela, c'est déjà faire un pas au-delà de la logique pure et embrasser une théorie non-conservative. La seconde étape consiste à introduire le prédicat de vérité proprement dit. Et, une fois accepté l'existence des objets as, la signification de « Vrai » est, quant à elle, effectivement bien fixée par des règles conservatives, c'est-à-dire conservatives sur notre nouvelle théorie qui n'est plus la logique pure, mais une théorie contenant des noms pour chaque énoncé du langage. Aussi, ce qu'il faut blâmer pour la non-conservativité des règles Vr-Intro/Vr-Elim, ce n'est pas la vérité, c'est tout simplement la syntaxe, c'est-à-dire l'acceptation de toute une collection d'objets constituée par les noms des énoncés de notre langage. Les règles pour « Vrai » satisfont donc « presque » l'harmonie globale. En quelque sorte, elles satisfont un critère de conservativité amendé : la vérité est une notion « quasi logique » ou « logico-syntaxique » dans la mesure où sa signification est fixée par des règles structurelles, satisfaisant l'harmonie locale et conservatives sur une théorie de la syntaxe du langage considéré. Elles ne permettent pas de prouver plus que ce qu'une théorie minimale de la syntaxe permet déjà d'établir. La vérité ne nous donne rien de plus que ce que l'on a dû préalablement accepter pour pouvoir formuler les règles qui en donnent la signification.

Par conséquent, s'il est exact que, d'après le critère inférentialiste de logicité, la notion déflationniste de vérité ne peut pas, à strictement parler, être qualifiée de logique, le fait qu'elle soit conservative sur tout contexte dans lequel on s'est déjà doté de noms permettant de désigner les énoncés de notre langage montre que cet échec n'est que relatif. Après tout, la thèse centrale du déflationnisme est que le prédicat de vérité est uniquement un outil de décitation augmentant le pouvoir expressif de notre langage, et que la notion de vérité n'a pas de contenu substantiel propre, ni de rôle explicatif important à jouer dans notre discours sur le monde. Sans soute faut-il, en toute rigueur,

qualifier la vérité de notion logico-syntaxique plutôt que strictement logique, mais ce glissement ne remet pas en cause l'intuition fondamentale des déflationnistes au sujet de la vérité. Le critère amendé d'harmonie globale rend justice à leurs idées sans toutefois lier la vérité à l'austère neutralité ontologique traditionnellement attribuée à la logique. C'est du moins à une conclusion de ce type qu'Henri Galinon parvient au terme de son analyse :

> La question de l'harmonie des règles va sans difficulté majeure. Il n'est pas difficile en effet de voir que ces règles satisfont les deux critères d'harmonie globale et locale que nous avons présentés dans la section précédente. Le critère de conservativité (en dépit des doutes que nous avons émis à son égard) est satisfait, comme le montre le résultat de conservativité des extensions minimales des théories par les *équivalences-T* que nous avons déjà mentionné, et en vertu de la correspondance entre règles-*Vr* et équivalences-T expliquée plus haut. Toutefois, ce que montre ce résultat est la conservativité sur le fond d'une théorie syntaxique donnée en avance, qui permette d'affirmer l'existence de certaines expressions, et peut-être les lois fondamentales de la formation de certaines classes d'expressions particulières. Ces règles ne sont pas conservatives sur la logique pure, puisque l'on peut en inférer l'existence de plusieurs objets. *Le critère de conservativité est un critère relatif, et l'harmonie globale de règles données est relative à un langage dans lequel les règles sont introduites.* Ce que l'on peut dire est que les règles pour la vérité satisfont le critère d'harmonie globale dans les contextes contenant un minimum de ressources syntaxiques. (Galinon, 2010, p. 327, nous soulignons)

Puis plus loin :

> Par conséquent, au vu de l'analyse précédente, *le prédicat de vérité est aussi logique qu'un prédicat s'appliquant à des énoncés peut l'être d'après les lumières de l'analyse inférentielle de sa signification* [...] (Galinon, 2010, p. 335, italiques de l'auteur)

Cependant, la non-conservativité n'était qu'un des aspects problématiques que nous avons relevés concernant la supposée logicité du prédicat de vérité déflationniste. L'autre problème concernait le caractère *purement* inférentiel de ses règles d'Introduction et d'Elimination.

2.3.2.2 Des règles minimales purement structurelles ?

Comme l'a souligné H. Hodes (2004), les règles d'Introduction et d'Elimination proposées pour « Vrai » ne sont pas à première vue *purement* structurelles. En effet, le passage de l'énoncé « A » à l'énoncé « Vr(a) », où « a » est un nom de l'expression « A » suppose que l'agent cognitif soit en mesure de déterminer la référence du terme « a » comme étant un certain énoncé du langage. La règle s'appuie donc crucialement sur une information sémantique. Or, c'était tout le sens de l'approche inférentialiste de la signification des constantes logiques que de proposer une caractérisation de celles-ci qui soit totalement indépendante de toute notion sémantique préalable. Le point central est que les expressions qui dénotent des constantes logiques peuvent être isolées et différenciées des autres expressions du langage à partir du rôle particulier que ces notions jouent dans nos raisonnements et nos pratiques déductives. À la suite de Gentzen, les partisans d'une approche preuve-théorique de la logicité ont défendu l'idée que les constantes logiques ne sont que les marques, ou les traces, de nos pratiques inférentielles fondamentales. [30] Elles désignent les « atomes » à partir desquels nous construisons nos raisonnements valides. Leur signification est donc caractérisable par un schéma de preuve, [31] qui expose le pas inférentiel minimal correspondant à la constante en question. [32]

De ce point de vue, la moindre considération sémantique parasitaire apparaissant dans la formulation des règles remet en cause la nature purement logique de l'expression considérée. Cela a en effet pour conséquence que la signification de l'expression considérée n'est pas entièrement ou, si l'on peut dire, « exhaustivement » donnée par une règle de preuve *purement* schématique. Or c'était justement là ce qui était censé caractériser les notions logiques. Dès lors, pour ce qui concerne les règles déflationnistes pour « Vrai », l'indispensable passage d'un énoncé « A » à un nom de cet énoncé, est loin d'être un

30. On peut songer ici à la fameuse et frappante formule de Wittgenstein :

> « Les signes des opérations logiques sont des signes de ponctuation. » Wittgenstein, 1922, 5.461, p 82

31. Dans un système de déduction naturelle, ce schéma de preuve est donné par des règles d'introduction et d'élimination (ou par un sous-ensemble de celles-ci).

32. Ce point est souligné de manière particulièrement claire par Peter Milne :

> [. . .] L'approche preuve-théorique affirme qu'une constante logique est définie par ses règles d'introduction et d'élimination ou par un sous-ensemble de celles-ci. Elle est preuve-théorique précisément en ce qu'elle refuse tout rôle aux considérations sémantiques dans la détermination de la signification des constantes logiques : c'est leur rôle dans les inférences qui fixe leur signification. (Milne, 1994)

problème bénin. Le problème posé par cet écart apparaît de manière particulièrement saillante lorsque la détermination de la référence de l'expression [33] à laquelle le prédicat de vérité s'applique est particulièrement opaque. [34]

Néanmoins, face à cette autre difficulté , on peut là encore proposer une réponse pour la défense des conceptions déflationnistes. Comme en écho à l'analyse du problème de l'harmonie globale, l'idée centrale va de nouveau consister à isoler ce qui relève proprement de la vérité et à le séparer de considérations ou ressources auxiliaires pour essayer de montrer que ce « coeur » des règles commandant le prédicat de vérité satisfait les conditions caractéristiques de la logicité.

Ainsi, réexaminant les règles *générales* données par Hodes pour le prédicat de vérité :

$Vrai$-Introduction :	$Vrai$-Elimination :	où « A » est un énoncé du langage considéré, et « a » un nom **quelconque** de cet énoncé,
Vr-In-générale $\dfrac{A}{Vr(a)}$	$\dfrac{Vr(a)}{A}$ Vr-El-générale	

Henri Galinon propose de distinguer des règles proprement minimales, formant une sous-classe incluse dans les règles générales, et qui suffisent à fixer la signification du prédicat de vérité, puis de les démêler clairement des ressources auxiliaires qui entrent en jeu dans les inférences du type de celles permises par les règles ci-dessus. Ces règles sont directement inspirée du la théorie quinienne de la décitation. Le point problématique, nous l'avons vu, est que la détermination de la référence du nom « a » constitue une connaissance de nature sémantique. Mais, nous dit Henri Galinon, pour chaque énoncé de notre langage, il existe un nom canonique : « *quelque chose comme ce que l'on obtient par la mise entre guillemets de cet énoncé* » [35] . Ce nom canonique est « *véritablement*

33. dénotant un énoncé, soit qu'elle le nomme directement, soit qu'elle en constitue une description définie.

34. Pour illustration, nous renvoyons aux exemples développés dans la section 2.3.1.

35. Pour être tout à fait exact, Henri Galinon propose deux « réductions » de ce type. Dans un premier temps, il suggère de prendre comme noms canoniques de nos énoncés des descriptions structurales —telles que celles développées par Tarski dans le *Wahrheitsbegriff*— obtenues à partir d'une théorie syntaxique minimale. Puis, poussant plus loin son analyse, il introduit les noms canoniques tels que nous les présentons ici. Cependant, le point d'arrivée, le plus important de notre point de vue, est bien celui indiqué : avec chaque énoncé de notre langage nous est immédiatement donné un nom canonique, sorte de nom propre que nous pouvons convoquer à volonté, sans devoir pour cela nous lancer dans une quelconque enquête ou entreprise de connaissance.

transparent » et le processus qui permet de passer d'un énoncé au nom canonique de ce dernier a un « *caractère d'immédiateté* » [36]. Le point crucial est que la spécification de la référence de ces noms canoniques est immédiate et ne fait pas appel à des ressources importantes, contrairement à ce qui peut arriver dans les cas d'application de la règle générale *à la* Hodes. Henri Galinon va même jusqu'à proposer de « *considérer ces noms d'énoncés comme donnés d'emblée avec le langage lui-même, comme des primitifs que je suis capable d'invoquer à volonté* [...] » [37]. Si l'on note «$_c$ A $_c$», les noms canoniques de nos énoncés, les règles « minimales » suivantes suffisent à caractériser la signification du prédicat de vérité :

Définition. *Règles Décitationnelles Minimales pour « Vr »* :

$$Vr\text{-In-Min}\ \frac{A}{Vr(\text{«}_c\ A\ _c\text{»})} \qquad \frac{Vr(\text{«}_c\ A_c\text{»})}{A}\ Vr\text{-El-Min}$$

où « «$_c$ A_c» » désigne le nom **canonique** de l'énoncé « A »

Il faut bien saisir la différence fondamentale de ces règles par rapport aux règles générales de Hodes. Ici, le passage de « A » au nom « «$_c$ A $_c$» » est censé être immédiat et transparent. L'obstacle « sémantique », le grain de sable qui remettait en cause la nature purement inférentielle des règles pour la vérité, est, si ce n'est entièrement levé, du moins grandement atténué, comme gommé. Les emplois plus généraux du prédicat de vérité, en particulier les règles de Hodes, doivent pouvoir être dérivés et justifiés à partir des seules règles *minimales* ci-dessus, avec le concours —si besoin est— d'autres ressources qui peuvent être considérables.

Ainsi, pour reprendre notre exemple précédent tiré de l'histoire napoléonienne [38], l'inférence suivante, où β est un nom de la dernière phrase prononcée par Napoléon lors de son allocution aux troupes à Embabèh,

$$\frac{\text{Vr}(\beta)}{\text{Du haut de ces pyramides quarante siècles vous contemplent}}\ Vr\text{-El-générale}$$

36. *cf.* Galinon, 2010, p. 332
37. *cf.* Galinon, 2010, p. 332
38. *cf.* section 2.3.1.

peut maintenant être décomposée en deux mécanisme distincts. Une première étape consiste à déterminer la référence de β et pour cela à établir que cette référence est identique à celle d'un de ces noms canoniques primitifs qui nous sont donnés immédiatement avec le langage lui-même. Puis, une fois cette identification réalisée, une deuxième étape consiste à appliquer la règle minimale d'Elimination :

$$\text{PSI}\ \dfrac{\dfrac{Vr(\beta) \qquad \beta = \text{«}_c\ \text{Du haut de ces pyramides quarante siècles vous contemplent}\ _c\text{»}}{Vr(\ \text{«}_c\ \text{Du haut de ces pyramides quarante siècles vous contemplent}\ _c\text{»})}\ Vr\text{-El-Min}}{\text{Du haut de ces pyramides quarante siècles vous contemplent}}$$

Cette décomposition met bien en lumière le rôle restreint que jouent les règles pour « Vr » dans cette déduction. Il est indéniable que la spécification de la référence de β (ou de la description définie : « La dernière phrase prononcée par Napoléon lors de son allocution aux troupes le 21 Juillet 1798 à Embabèh ») mobilise un nombre important de connaissances qui vont bien a-delà de la simple logique, ou de la simple définition inférentielle de « Vr ». Mais, il semble clair que le prédicat de vérité n'a pas de rôle à jouer en cela. Il n'apparaît que dans un deuxième temps, après que la référence de β a été identifiée à celle d'un nom canonique de l'énoncé effectivement prononcé par Napoléon. Une fois clairement séparé ce qui relève en propre de la vérité, le rôle joué par le prédicat de vérité dans nos pratiques déductives, et par conséquent sa signification selon le paradigme inférentialiste, peuvent être fixés au moyen des règles plus modestes que sont les règles minimales.

Accepter cette réduction du rôle inférentiel du prédicat de vérité, « *c'est semble-t-il,* [nous dit Henri Galinon], *faire droit à l'idée que l'inférence de la vérité d'un énoncé que je me* présente *comme objet de pensée, à l'affirmation de cet énoncé lui-même,* ne mobilise aucune véritable syntaxe [39]*, c'est faire comme si l'intimité de notre commerce avec nos énoncés était telle que nous aurions un* nom propre *de chacun de ceux que nous utilisons.* » Dès lors, les ressources nécessaires pour rendre compte de la signification du prédicat de vérité sont beaucoup plus modestes que ce que les règles générales proposées par Hodes pouvaient laisser croire. Outre les règles d'introduction et d'élimination minimales, on doit simplement se donner les noms canoniques « «$_c$ A $_c$» » pour tous les énoncés « A » de notre langage et des axiomes habituels de l'égalité du type $\vdash t = t$, le principe de substitution des identiques (PSI) et les axiomes particulièrement élémentaires suivants :

39. Nous soulignons.

Si A, B sont des énoncés de notre langage et que $A \neq B$:

$$\vdash \text{«}_c\ A\ _c\text{»} \neq \text{«}_c\ B\ _c\text{»}$$ [40]

Ces ressources ne sont pas purement logiques[41], néanmoins, « *la théorie des objets nécessaires pour rendre compte des inférences aléthiques est* particulièrement *faible* »[42].

On ne peut pas totalement contourner le problème posé par le fait qu'il est nécessaire d'appliquer le prédicat de vérité à des *noms* d'énoncés, et que d'une part ces objets nouveaux charrient un engagement ontologique, et que, d'autre part, la référence de ces noms peut paraître problématique à établir. Néanmoins, en posant l'existence de nom canoniques pour chacun des énoncés de notre langage, Henri Galinon parvient à argumenter en faveur du caractère « quasi » purement inférentiel de règles minimales censées suffire à fixer la signification de « Vrai ». Ces noms canoniques, sorte de noms propres de nos énoncés, sont des *primitifs* qui nous sont *immédiatement donnés* et présentent de façon *transparente* l'énoncé qu'ils dénotent : le passage d'un énoncé « A » à l'énoncé « Vr(«$_c$ A $_c$») » se fait donc sans difficultés, sans mobiliser aucune véritable syntaxe ou connaissance sémantique, qui mettraient en cause de façon significative la nature purement inférentielle de ce pas déductif.

Ceci permet à Henri Galinon de conclure qu' « *au vu de l'analyse précédente, le prédicat de vérité est aussi logique qu'un prédicat s'appliquant à des énoncés peut l'être d'après les lumières de l'analyse inférentielle de sa signification* [...] »[43]. Pas totalement logique, mais presque.

2.4 Critique de l'argument

Nous voudrions à présent évaluer la force de cette analyse et proposer quelques critiques et observations.

2.4.1 Portée de l'argument

La première remarque qu'il nous paraît indispensable de formuler est à la fois la plus évidente et sans doute la plus importante. Elle concerne la portée de l'analyse développée.

40. *Cf.* Galinon (2010, p. 334).
41. En particulier, elles permettent de prouver l'existence de plusieurs objets (autant, qu'il y a de noms canoniques introduits). Ce qui nous amène au-delà de la logique pure.
42. *Cf.* Galinon (2010, p. 334).
43. Galinon (2010, p. 335).

Nous avons déjà évoqué très brièvement ce problème. Mais il nous semble important d'en dire un peu plus. Un argument pour la logicité de la vérité s'appuyant sur des règles du type de celles avancées ici [44] ne peut être convaincant que si l'on accepte au départ, avant même d'en tester la logicité, que les règles structurelles proposées disent bien *tout* ce qu'il y a à dire sur la vérité, qu'elles fixent bien *entièrement* la signification du prédicat « vrai », ou du moins sa signification inférentielle. Une telle position s'apparente à accepter déjà les thèses déflationnistes. Si l'on n'est pas d'emblée un peu convaincu que la vérité est une notion « purement décitationnelle » dont il s'agirait simplement d'expliquer le rôle expressif au sein de notre langage, rôle expressif qui suffirait à en expliquer l'importance et l'indispensabilité, l'argument est donc sans force : comme nous l'avons déjà remarqué, les partisans d'une théorie plus riche, ou plus « substantielle », de la vérité refuseront cette réduction de la signification de « vrai » à de telles règles. [45]

Pour autant, le déflationnisme en matière de vérité ne se limite pas à l'affirmation que les règles Vr-Intro/Elim (ou, dans la version plus habituelle : que l'ensemble des T-équivalences) fixent la signification du prédicat de vérité. À cette position concernant la bonne axiomatisation de la notion, s'ajoutent des thèses plus proprement philosophiques quant à la nature (ou l'absence de nature) de la vérité. Simple outil expressif, « vrai » ne désignerait pas une notion substantielle ; la vérité n'aurait pas de rôle explicatif à jouer dans nos théories ; elle serait « neutre » d'un point de vue tant épistémologique que métaphysique ; enfin, et c'est le point qui nous intéresse ici, elle serait une notion logique. Les analyses précédentes ne sont donc pas tant un argument en faveur de la logicité de la vérité *in abstracto* ou en toute généralité. Il convient plutôt d'y voir une tentative d'évaluer la cohérence des positions déflationnistes. Il s'agit plus, en quelque sorte, d'un raisonnement sous hypothèse : *si* l'on accepte que les règles déflationnistes fixent entièrement la signification du prédicat « vrai », a-t-on de bonne raisons de penser que ce prédicat est une sorte de constante logique ? [46]

En somme, les déflationnistes courent un véritable risque pour une maigre victoire potentielle : si, partant d'un prédicat caractérisé par des règles telles que celles présentées ici, on parvenait à la conclusion que ce prédicat ne satisfait pas nos conditions de logicité, on aurait mis au jour une grave tension au sein des thèses déflationnistes. Si, à l'inverse,

44. Que ce soit dans la version *générale* de H. Hodes, ou que ce soit les règles *minimales* présentées par H. Galinon

45. Indépendamment de la question de savoir si de telles règles peuvent être qualifiées de logiques.

46. C'est d'ailleurs ainsi qu'Henri Galinon présente son analyse (*cf.* Galinon (2010, p. 309-336). Mais il ne nous semble pas inutile d'insister sur ce point.

on parvenait à la conclusion qu'un tel prédicat est bien logique, il serait toujours temps, pour les défenseurs d'une vision substantielle de la vérité qui récusent toute idée d'une vérité quasi logique, de remettre en cause le point de départ selon lequel les règles proposées disent bien *tout* ce qu'il y a à dire sur cette notion.

Ceci étant, et gardant présent à l'esprit le fait que nous raisonnons sous hypothèse, force est de constater que, même si l'on admet comme hypothèse de départ que la signification du prédicat de vérité est bien exhaustivement caractérisée par des règles du type Vr-Intro/Vr-Elim, la logicité de ces règles ne va pas sans poser un certain nombre de difficultés. Sur les trois conditions qui composent le critère de logicité que nous avons proposées ici à la suite de Dummett et Prawitz [47], les règles générales pour « Vr » n'en satisfont à strictement parler qu'un seul : l'harmonie locale, ou critère de réduction. Comme nous l'avons vu, les règles pour « vrai » ne valident pas le critère d'harmonie globale, pas plus qu'elles ne sont, à strictement parler, purement structurelles. Malgré ces obstacles, Henri Galinon a proposé une défense de la position déflationniste, qui commence par une réduction de la signification de « vrai » aux seules règles minimales. Cette défense nécessite certains aménagements du critère de logicité inférentialiste proposé. La conservativité n'est vérifiée que sur une théorie syntaxique. Et les règles sont « presque » purement structurelles, mais seulement si l'on accepte de postuler l'existence de noms canoniques pris comme primitifs. Dès lors, la crédibilité du raisonnement déployé dépend crucialement de la mesure dans laquelle ces aménagements apparaissent justifiés ou acceptables au regard des approches inférentielles de la logicité. Nous allons donc examiner un peu plus précisément les règles *décitationnelles minimales* et formuler certaines objections concernant la façon dont elles sont censées justifier la nature quasi logique du prédicat de vérité.

2.4.2 Retour sur les Règles *Minimales*

Comme nous l'avons expliqué, le but qui sous-tend l'appel aux règles minimales est de répondre au problème posé par le fait que les règles pour « vrai », dans la mesure où elles commandent d'introduire des noms d'énoncés, ne sont pas purement structurelles. Henri Galinon (Galinon (2010)) propose de « factoriser » les mécanismes qui relèvent de ce qu'il appelle « la syntaxe » —et qui comprennent, sans doute, les descriptions définies dénotant un énoncé, les outils permettant de donner une description structurelle des

47. Et d'autres, par exemple : Tennant, 1987, 1997, Milne, 1994,...

énoncés, ou bien les théories de la concatenation permettant de rendre compte de leur mise entre guillemets, voire, peut-être, dans les contextes d'arithmétique formalisée, un codage gödelien— afin de les séparer de règles qui doivent suffire à donner la signification en propre du prédicat de vérité. En posant l'existence de noms canoniques, notés «$_c$ A $_c$», il introduit les règles minimales dans lesquelles le passage de « A » à « Vr(«$_c$ A $_c$») » [48] est censé se faire de manière transparente et immédiate, ce qui les rapproche de règles purement structurelles. Il est indéniable que par leur forme graphique, les règles minimales semblent gagner en « structuralité » : tout comme dans les règles pour les constantes logiques standard, c'est bien les mêmes variables désignant des énoncés qui apparaissent de chaque côté de la barre d'inférence, [49] à ceci près que certaines d'entre elles sont flanquées de guillemets. En outre, lorsqu'on nous présente une instance de ces règles telle que :

$$\frac{\text{«}_c \text{ La neige est blanche } {}_c\text{» est vrai}}{\text{La neige est blanche}}\ Vr\text{-El-Min}$$

nous sommes évidemment envahis par un sentiment de trivialité. Cependant, malgré les apparences, cette construction ne nous semble pas répondre de manière adéquate au problème posé. Elle ne nous dit rien, ou presque rien, sur la nature de ces noms canoniques. Seules sont postulées certaines de leurs propriétés : ils sont « immédiats », « transparents », « primitifs ». Quant à la manière dont ils sont obtenus, nous devons nous contenter de savoir qu'ils sont « quelque chose comme ce que l'on obtient par une mise entre guillemets ». [50]

2.4.2.1 Quel statut pour les noms canoniques ?

Le statut de ces noms canoniques, tels qu'ils sont introduits par Henri Galinon, nous apparaît en fait un peu ambigu. Dans un premier temps, il affirme que :

> ... pour expliquer notre usage inférentiel du prédicat de vérité, ***il n'est pas besoin de supposer que nous disposions véritablement d'une théorie syntaxique***, avec ces notions d'expressions et ses lois de concaténation. Il y a, me semble-t-il, un sens dans lequel chaque énoncé possède un nom canonique véritablement transparent : quelque chose comme ce que l'on obtient par la mise entre guillemets de cet énoncé. (Galinon, 2010, p. 332, nous soulignons)

48. ... et inversement.
49. Contrairement au cas des règles générales ou a désigne un terme quelconque dénotant l'énoncé A.
50. Mais, c'est déjà beaucoup puisque ce sont des primitifs.

Mais, quelques lignes plus tard, lorsqu'il s'agit de décrire un peu plus précisément les règles minimales, il a lui-même délibérement recours à une telle « théorie syntaxique », munie de lois de concaténation et de règles de formation d'énoncés, ce qui lui permet de construire explicitement des (notations pour les) noms canoniques. La règle minimale d'introduction se trouve alors ainsi formulée :

> Pour tout ensemble fini σ d'énoncés du langage, pour tout énoncé x de ce langage ne contenant pas le prédicat de vérité,
>
> si $\sigma \vdash x$ alors $\sigma \vdash \overline{Vr} * \overline{(} * \overline{\ulcorner} * x * \overline{\urcorner} * \overline{)}$. (Galinon, 2010, p. 333)) [51]

Dès lors, on est conduit à s'interroger. Les noms canoniques doivent-ils, ainsi que le suggère le premier extrait ci-dessus, être pris comme des primitifs laissés inanalysés au motif qu'ils seraient immédiatement présents à l'esprit de l'agent coginitif, sans que celui-ci doive disposer d'une véritable « théorie syntaxique » permettant d'indiquer comment ils acquièrent leur référence et de clarifier les lois qui gouvernent leur usage [52] ? Si tel est le cas, peut-être pouvons-nous alors considérer que les noms canoniques sont donnés d'emblée dans la formulation des règles minimales, sans qu'il soit nécessaire d'en dire plus à leur sujet. *(Option 1.)*

Ou faut-il au contraire expliquer comment ces noms d'énoncés sont formés et comment nous en venons à maîtriser leurs significations, ce qui rend nécessaire d'en donner une théorie explicite ? *(Option 2.)* Bien sûr, le danger est alors qu'une telle théorie dépassera largement les seules ressources de la logique. Nous aurons alors besoin d'une « théorie syntaxique » [53], et peut-être d'autres choses encore. Si l'on cherche à « factoriser » la syntaxe pour se concentrer sur des règles minimales aussi « logiques » que possible, il est tentant de privilégier la première option.

Mais il y a peut-être encore une troisième possibilité : plutôt que de postuler l'existence de noms canoniques inanalysables, ou d'en donner une théorie explicite, peut-être pourrions-nous *traiter* les noms canoniques *comme* des primitifs dans la formulation des règles minimales, tout en admettant que leur nature pourra être élucidée *par ailleurs*, au

51. où $*$ est un opérateur de concaténation et ou $\overline{s}$ est un nom pour le symbole s.

52. Dans ce cas, la mise entre guillemets : « la neige est blanche », n'est qu'une façon de *noter* ce nom primitif. Elle ne nous dit rien sur la manière dont on l'obtient à partir de l'énoncé « la neige est blanche ». On aurait tout aussi bien pu noter ce nom : **la neige est blanche**, ou encore : $\overline{\text{la neige est blanche}}$, ou plus facétieusement : Toto ...

53. Et encore, cette dernière épithète est peut-être mal choisie puisque ce qu'il s'agit d'expliquer ici est la *référence* de ces noms canoniques.

moyen d'une théorie indépendante.[54] Simplement, l'analyse du processus de nominalisation (canonique) doit être soigneusement séparée de la formulation des règles minimales. *(Option 3.)* Toutefois, dans une tentative d'élucidation de la signification du prédicat de vérité, une telle mise à l'écart du processus de nominalisation est-elle possible ? Est-elle justifiée ? Si la vérité est avant tout (uniquement ?) un outil de décitation ou de dénominalisation, une étude de sa nature plus ou moins logique peut-elle vraiment faire l'économie d'une analyse des phénomènes de citation ou de nominalisation qui opèrent derrière les noms canoniques.

Il nous semble que non et que, quelle que soit l'option choisie, la construction proposée ne permet pas de réellement réconcilier la notion déflationniste de vérité avec le critère inférentialiste de logicité, tel qu'il a été développé par Dummett et Prawitz. Essayons de voir pourquoi.

2.4.2.2 Des noms canoniques primitifs ?

Le point crucial d'achoppement, ce qui « empêche » les règles pour la vérité d'être purement structurelles, est la nécessaire présence d'un nom d'énoncé auquel s'applique le prédicat de vérité et dont la *référence* doit être *connue* pour qu'on puisse appliquer une règle d'introduction ou d'élimination. L'hypothèse de l'existence de noms canoniques ne nous semble pas lever cette difficulté. Le fait que, par hypothèse, la référence des noms canoniques soit triviale et immédiatement visible ne change pas sa nature. Et, c'est donc bien une information à caractère *sémantique* qui apparaît crucialement dans la manipulation du prédicat de vérité. Or, insistons-y, du point de vue de l'approche inférentielle de la logicité, la question n'est pas celle d'une différence de degré : les règles sont-elles « plus ou moins » purement structurelles ? La question est celle d'une d'une différence de nature : les règles sont-elles ou ne sont-elles pas purement structurelles ? Plus spécifiquement, dépendent-elles ou non du contenu sémantique des objets qui y apparaissent ? Qu'elles en dépendent un peu ou beaucoup, de manière plus ou moins patente, et que la connaissance de ce contenu sémantique soit elle-même triviale, immédiate, donnée *a priori*, innée, ou au contraire complexe et difficile à établir, ne change rien à l'affaire.

Certes, le « déficit d'inférentialité » des règles d'introduction et d'élimination est plus

54. Si nous l'avons bien compris, c'est sans doute cette troisième voie que propose Henri Galinon. *Cf.* Galinon, 2010.

criant dans une inférence du type :

(α)

$$\frac{\text{La dernière phrase prononcée par Napoléon lors de son allocution à Embabèh est vraie}}{\text{Du haut de ces pyramides quarante siècles vous contemplent}}\ Vr\text{-El-générale}$$

que dans une application de la règle d'Elimination Minimale :

(β)

$$\frac{\text{«}_c\text{ La neige est blanche }_c\text{» est vrai}}{\text{La neige est blanche}}\ Vr\text{-El-Min}$$

Dans la première inférence, la référence de la description définie sur laquelle porte le prédicat de vérité est loin d'être évidente. Elle est particulèrement « opaque », ce qui renforce le sentiment qu'une connaissance de nature sémantique —à savoir celle de l'objet dénoté par cette description définie— entre en jeu dans la manipulation du prédicat de vérité. Dans le second cas en revanche, le nom canonique, noté au moyen d'une mise entre guillemets de l'énoncé, semble porter sa référence sur ses épaules, si bien que le passage d'une ligne à l'autre s'effectue, du moins en apparence, de manière directe et « transparente ».

Mais cette impression nous paraît trompeuse. Ne sommes nous pas victimes d'une sorte d'illusion graphique ? Ayant noté[55] « «$_c$ A $_c$» » le nom canonique de « A », nous avons le sentiment que c'est presque la même chose, à une légère variation de catégorie grammaticale près, qui apparaît de chaque côté de notre barre d'inférence et que l'inférence formalisée par cette règle est triviale. Mais en fait, une connaissance implicite de la référence de « «$_c$ A $_c$» » est bel et bien indispensable pour pouvoir apprendre, maîtriser et appliquer les règles. Pour illustrer ceci, plaçons-nous dans une situation d'apprentissage. Puisque saisir la signification de « vrai » consiste simplement à maîtriser certaines règles, supposons que nous voulions enseigner la vérité à un agent cognitif qui serait totalement ignorant en cette matière. Supposons, par exemple, que nous voulions programmer un ordinateur[56] pour qu'il puisse employer correctement un prédicat déflationniste de vérité gouverné par les règles *minimales*. Que devrions-nous inscrire dans son programme ? À supposer que le programme maîtrise déjà un langage de base $\mathcal{L}$

55. ce qui, à soi seul, ne nous explique en rien comment un tel nom peut être obtenu, ni comment nous sommes capables de l'employer correctement.

56. Le choix de cet exemple informatique n'est pas fortuit : les partisans de l'approche inférentielle de la logicité insistent particulièrement sur la faisabilité effective, *i.e.* la calculabilité, des règles.

dans lequel la notion de vérité n'apparaît pas —et peut-être aussi les règles gouvernant les constantes de la logique du première ordre— il nous faudrait, pour chaque énoncé que la machine peut être amenée à manipuler, inscrire dans sa base données un nom canonique de cet énoncé. Ceci constitue un préalable indispensable à la programmation des règles d'introduction et d'élimination minimales. De plus, si le programme peut être amené à former et à manier un ensemble potentiellement infini d'énoncés (par exemple, au moyen de règles de formation syntaxique) [57], la simple donnée de noms canoniques par énumération des énoncés ne saurait suffire. Il nous faudra plutôt fournir à notre machine toute une théorie de la syntaxe [58] des noms d'énoncés lui permettant de construire un nom canonique pour chaque énoncé qu'elle pourrait avoir à manipuler. Quoi qu'il en soit, cela constitue une masse d'information non négligeable.

Dans le cas des locuteurs humains, postuler l'existence de noms canoniques primitifs qui seraient immédiatement disponibles et transparents et ne mobiliseraient aucune « véritable syntaxe » [59], ou pour le dire autrement, postuler l'existence d'un mécanisme « natif » de nominalisation des énoncés, ne fait, nous semble-t-il, que masquer l'importance et la complexité du phénomène. Ce mécanisme est peut-être inné, ou acquis solidairement lors de l'apprentissage d'une langue. Mais cela n'en fait pas pour autant un mécanisme purement logique. Bien au contraire, il sera probablement aussi complexe et réclamera des ressources aussi riches que le genre de programmes qu'il nous a fallu entrer dans notre machine lors de l'exemple précédent. La complexité de ce mécanisme apparaîtrra d'autant plus clairement si l'on garde présent à l'esprit que, de fait, nous, humains, devons pouvoir nommer (canoniquement) un ensemble potentiellement infini d'énoncés. Que cette somme d'informations nous paraisse familière et triviale ne change rien ni à sa nature, ni au fait que l'accès à cette information est un préalable indispensable à l'application des règles déflationnistes pour « vrai ». [60]

57. Notez que tel sera le cas, par exemple, si la machine maîtrise les constantes logiques standard : à partir de A déduire $A \vee A$, puis $(A \vee A) \vee A$, puis *etc.*

58. Ou devrions-nous dire une théorie *syntaxico-sémantique* permettant non seulement d'assembler syntaxiquement les noms canoniques mais donnant aussi leurs références ?

59. Nous reprenons ici la terminologie d'Henri Galinon.

60. D'ailleurs, *a contrario*, aux yeux d'un spécialiste de l'histoire napoléonienne ayant passé de nombreuses années dans la compagnie de l'Empereur, la référence de l'expression « la dernière phrase prononcée par Napoléon lors de son allocution aux troupes à Embabèh le 21 juillet 1798 » est peut-être tout aussi transparente, immédiate, et évidente que celle du nom canonique « « Du haut de ses pyramides quarante siècles vous contemplent » ». De même, la façon dont un mathématicien féru de théorie des ensembles « nominalise intérieurement » un certain énoncé parfois sujet à controverse lorsqu'il discute avec ses collègues, s'appuiera vraisemblablement plus sur la locution « L'Axiome du choix » plutôt que sur le nom canonique « « $\forall A \exists F : A \backslash \{\emptyset\} \longrightarrow \bigcup A \forall x \in A \backslash \{\emptyset\}(F(x) \in x)$ » ». Ainsi, l'impression de

Plus généralement, des règles minimales appuyées sur l'existence de noms canoniques considérés comme primitifs et immédiatement donnés s'accordent mal avec les principes de l'analyse preuve-théorique de la logicité. Comme nous l'avons expliqué précédemment, l'idée fondamentale qui sous-tend toute l'approche inférentielle des constantes logiques est que leurs significations peuvent être caractérisées par des règles d'inférence particulières. Mais, d'une part, ces règles sont censées donner *exhaustivement* la signification des constantes considérées, au sens où il faut et il suffit de maîtriser les règles pour saisir la signification de la constante. Et, d'autre part, les règles sont supposées être *publiquement observables.*[61] Sous ces deux aspects au moins, l'idée que, dans le cadre d'une évaluation de la logicité des règles pour « vrai », les noms canoniques puissent être considérés comme primitifs, sans qu'il soit besoin d'éclaircir leur nature, est inacceptable.

Concernant l'exhaustivité : si les règles sont censées fixer entièrement la signification de la notion qu'elles caractérisent, si elles disent tout ce qu'il y a à en dire, alors elles doivent mettre au jour *tous* les mécanismes qui président à la saisie de cette signification. C'est ce que montrent les situations d'apprentissage : les règles doivent incorporer toutes les connaissances nécessaires à l'emploi du concept qu'elles définissent, puisque comprendre les règles suffit à saisir pleinement la signification du concept. Dès lors, lorsqu'il s'agira d'évaluer la logicité des règles (ou la substantialité de la notion qu'elles définissent), il faudra tenir compte de tous les ingrédients qui y apparaissent. Qualifier certains d'entre eux de primitifs, ne peut être au mieux qu'une étape provisoire qui ne dispense pas de les prendre en considération.

Une comparaison avec la forme habituelle des définitions en logique permet d'illustrer ce point. Dans une définition classique, le *definiendum* est défini au moyen d'un équivalence logique le reliant à un *definiens*. Et, si l'on veut juger de la logicité ou de la « substantialité » d'un *definiendum*, il est indispensable de prendre en compte la nature de tous les termes qui apparaissent dans le *definiens*, quand bien même certains de ces termes seraient considérés comme des primitifs, soit qu'on les considèrent comme trop fondamentaux pour être à leur tour analysés, soit qu'on en remette l'analyse à plus tard. Tel *definiendum* sera logique si seuls des termes et des relations logiques apparaissent dans le *definiens* correspondant ; tel concept pourra être qualifié d'arithmétique si son *definiens* ne contient que des notions arithmétiques ; telle notion de psychologie popu-

triviale familiarité et de transparence est peut-être toute relative.

61. D'après les tenants de l'approche inférentialiste, ces deux caractéristiques sont en particulier nécessaires pour que l'on puisse apprendre la signification d'une constante.

laire sera en fait une notion physique si on parvient à la relier à un *definiens* construit à partir des seules ressources de la physique, *etc.* Le cas de règles structurelles fixant la signification d'une expression n'est pas si différent.

Voici un exemple qui permet d'illustrer ce que nous avons en tête. Supposons que, suite aux progrès de l'imagerie médicale, nous soyions en mesure d'identifier la douleur à un certain état particulier du système nerveux. Nous pourrions alors donner une définition de la propriété « avoir mal » :

$$x \text{ a mal ssi } x \text{ est dans telle et telle configuration neurologique.}$$

Si l'on note « Rx » la propriété « être dans telle et telle configuration neurologique », on a donc

$$x \text{ a mal} \leftrightarrow Rx.$$

Mais, supposons que je m'interroge sur la logicité éventuelle de notre concept de douleur, et que lors de l'étude de cette question je m'autorise à considérer R comme un élément primitif de ma théorie.[62] Je pourrais alors développer un curieux argument :
Les règles suivantes

"Définition". *Règles Minimales pour « a mal »* :

$$\textit{Mal}\text{-Intro}\ \frac{Rx}{x \text{ a mal}} \qquad\qquad \frac{x \text{ a mal}}{Rx}\ \textit{Mal}\text{-Elim}$$

qui ne font que traduire l'équivalence élémentaire de notre définition par des règles couchées en déduction naturelle, suffisent à fixer la siginifcation inférentielle du prédicat « a mal ». Elles sont « presque » purement structurelles du moment que l'on accepte de considérer R comme primitive et immédiatement donnée. Au passage, remarquons qu'elles auront aussi le bon goût d'être conservatives sur toute théorie contenant R —ce qui résout le problème de l'harmonie globale, du moins si l'on prend comme condition de logicité un critère relatif de conservativité. En outre, on ne saurait exiger d'une théorie de la douleur qu'elle rende compte du fonctionnement du système nerveux. C'est bien

62. Peut-être de manière seulement provisoire, en attendant d'analyser *par ailleurs* la propriété R, au moyen d'une théorie *indépendante et clairement séparée* des règles pour « a mal ».

à une théorie indépendante[63] s'appuyant sur des ressources pouvant être considérables, que nous devrons nous adresser pour éclaircir la nature de R. Mais, cela ne doit pas nous retenir de définir par ailleurs le prédicat « a mal » de manière quasi logique.

Bien sûr, un tel argument en faveur de la logicité de la douleur n'est guère convaincant. Et, il semble clair qu'une évaluation de la logicité des règles fixant la signification de l'expression « a mal » doit prendre en compte tous les éléments, primitifs ou non, qui entrent en jeu dans la maîtrise de leur usage.

Pourquoi le cas de la vérité déflationniste serait-il différent ? Certes, en la circonstance, nous n'avons pas affaire à une définition, classique dans sa forme, de la notion de vérité. Les théories déflationnistes du concept de vérité se présentent le plus souvent sous l'aspect d'une axiomatisation donnée par la collection des instances du schéma-T.[64] Par conséquent, comme nous ne partons pas d'une unique équivalence reliant *definiens* et *definiendum* mais de la collection infinie des T-équivalences, lorsque nous reformulons notre théorie dans un cadre de déduction naturelle, nous obtenons comme pendant inférentiel de cette axiomatisation non pas une unique paire de règles d'Introduction et d'Elimination mais plutôt une collection infinie de telles règles (une par T-équivalence, autrement dit une pour chaque nom d'énoncé). Si l'on préfère, on peut considérer les règles minimales comme un schéma de règles[65] qui expose comment obtenir une instance de la règle en s'appuyant sur une connaissance de la référence du nom d'énoncé qui y émerge. Le résultat en est que, contrairement au pseudo argument sur la douleur, nous employons comme primitifs dans la formulation des règles non pas un unique *definiens*, mais une infinité de noms canoniques. Néanmoins en quoi cela nous dispenserait-il d'éclaircir la nature de ces primitifs, de théoriser les mécanismes qui les gouvernent et de tenir compte du rôle qu'ils jouent lorsque nous tentons d'évaluer la logicité des règles pour « vrai » ?

À partir des règles générales de Hodes, il est certainement possible d'isoler un sous-ensemble d'usages proprement définitionnels de la vérité.[66] Une fois cette restriction réalisée, on peut voir dans l'inférence (α) ci-dessus[67] un emploi dérivé du prédicat de

63. En l'occurrence la neurobiologie

64. Cette axiomatisation est parfois qualifiée de définition implicite, en dépit du théorème de Beth. Pour une discussion sur ce point voir Bays (2009) et Ketland (2009).

65. qui correspond bien sûr au schéma-T de Tarski.

66. Mais nous allons voir que, dans le cas précis de la vérité déflationniste, il y a des raisons supplémentaires de mettre en doute la pertinence d'une telle « factorisation de la syntaxe » aboutissant à la restriction aux règle minimales (*cf. infra* **??**).

67. ou plus généralement dans les instances des règles générales qui ne sont pas des instances des règles

vérité qui s'appuie sur des ressources autres que celles relevant de la pure vérité, et pour lequel on ne saurait demander des comptes à la vérité *tout court*. Mais même si on admet que les usages représentés par les règles minimales [68] suffisent à fixer la signification du prédicat de vérité, il n'en demeure pas moins que ces règles elles-mêmes s'appuient sur une connaissance de la référence des noms canoniques qui y figurent. Les règles minimales exhibent peut-être toutes les connaissances nécessaires à la maîtrise de la signification du prédicat « vrai » mais cela ne signifie pas qu'elles en constituent l'analyse finale. Au contraire, nous sommes en droit d'attendre une élucidation des éléments primitifs qui apparaissent dans leur formulation. Comme dans le cas de définitions classiques, si en s'appuyant sur ces connaissances préalables, les règles minimales suffisent à donner une définition inférentielle de la signification du prédicat « vrai », lorsqu'il s'agira de juger de leur logicité, il faudra prendre en compte tous leurs constituants. [69]

L'argument que nous venons de développer s'applique évidemment à la première option que nous avons distinguée à propos du statut des noms canoniques et qui consiste à les prendre comme des primitifs immédiatement donnés et ne réclamant pas d'analyse supplémentaire. Mais il est important de remarquer que cet argument s'applique également à la troisième voie qui entend *prendre* les noms canoniques comme des primitifs dans la *formulation* des règles minimales, quitte à *analyser ultérieurement* leurs propriétés au moyen d'une théorie indépendante, et tirer profit de cet isolement des règles minimales par rapport à une théorie des noms canoniques pour en défendre la logicité.

minimales

68. On pourrait les appeler usages purement décitationnels

69. Par anticipation, nous voudrions répondre à une critique possible à laquelle notre analyse pourrait en apparence se prêter. D'après ce que nous avons défendu, il importe de prendre en compte tous les ingrédients d'une règle lorsqu'on veut en évaluer la logicité. Mais ne court-on pas alors le risque d'une régression à l'infini ? Supposons que nous considérions des règles censées fixer la signification d'une expression : pour être formulées, ces règles devront s'appuyer sur un certain nombre d'éléments (par exemple des meta-variables représentant des énoncés, une barre d'inférence tracée entre des hypothèses et une conclusion, et peut-être d'autres choses encore ...). Ces éléments devront être à leur tour analysés, et il faudra expliquer leur signification, par exemple au moyen d'autres règles. Mais ce processus doit avoir une fin, sous peine d'entrer dans un mouvement de régression infinie (ou de cercle vicieux). Dès lors, s'il est indispensable, y compris pour l'analyse inférentielle de la signification des constantes logiques, d'avoir « un point de départ », pourquoi ne pas inclure les nom canoniques parmi ces primitifs sur lesquels il faut bien s'appuyer pour commencer l'analyse ? Ce contre-argument s'appuie en fait sur un compréhension erronée de la démarche inférentielle. Pour les inférentialistes, tels que Dummett ou Prawitz, le propre des notions logiques est justement que les seuls points de départ acceptables pour leur analyse sont les notion d'inférence et de règles. Ces points de départ nous sont fournis par le cadre de la déduction naturelle (ou du calcul des séquents) qui permet de noter nos inférences. Mais en dehors de cela, tout autre ingrédient est prohibé. C'est, en somme, le coeur de l'approche inférentielle de la logicité que de considérer que tout autre élément —et au premier chef toute autre connaissance sémantique— intervenant dans la formulation des règles nous fait aussitôt sortir de la simple logique.

Même si pour des raisons méthodologiques [70] nous décidons de traiter à part la question de la nature des noms canoniques et de remettre à plus tard l'étude de leurs mécanismes, dans la mesure où ils apparaissent dans les règles minimales et où la connaissance de leurs références est indispensable à une maîtrise des règles minimales, une évaluation de la logicité de ces règles doit prendre aussi en compte les mécanismes qui gouvernent l'usage des noms canoniques [71]. Or, pour préciser la manière dont nous saisissons la référence de ces noms canoniques, nous devrons certainement introduire des ressources qui dépassent les règles minimales et le cadre de la logique. Autrement dit, le fait qu'on qualifie éventuellement de primitifs les noms canoniques et qu'on les traite comme tels dans la formulation des règles minimales ne doit pas servir à masquer leur importance.

Si l'on se tourne à présent vers l'impératif de publicité des règles, il nous semble évident que celui-ci est également mis à mal par l'hypothèse de noms canoniques primitifs, du moins si on entend par là qu'ils sont donnés d'emblée et immédiatement disponibles pour l'agent cognitif, sans qu'il soit besoin de fournir une explication supplémentaire. [72] Une telle conception semble en effet sanctionner un usage privé des noms canoniques. Quelle est cette faculté primitive dont tous les locuteurs sont censés être uniformément dotés et qui leur permet de nommer les énoncés et de passer immédiatement d'un nom d'énoncé à cet énoncé lui-même ou inversement ? Quelle garantie avons-nous que cette « lumière naturelle » soit partagée par tous de la même manière et que nous en fassions tous le même emploi ? Rappelons que pour les partisans de l'approche inférentialiste, le langage est un bien public et la signification de chaque expression doit pouvoir être saisie en commun par tous les locuteurs compétents au travers des usages publiquement observables qui en sont faits. Au regard de ce paradigme, il est donc tout aussi impératif de clarifier dans le cas des noms canoniques comment nous pouvons saisir leur signification, notamment leurs références, à partir des usages publics que nous en faisons. Aux yeux d'un inférentialiste, une règle qui repose sur une « faculté » non analysée ou « primitive » de l'agent cognitif, lui permettant de discerner la valeur sémantique d'un objet, ressemble donc fort à un constat d'échec dans la tentative d'élucidation de

70. . . . ou idéologiques.

71. Tout comme une évaluation de la logicité de « a mal » devait prendre en compte la théorie sous-jacente nécessaire pour éclaircir la nature du terme « R », bien que celui-ci soit pris comme primitif dans la définition.

72. Cette critique s'adresse donc directement avant tout à la première option que nous avons distinguée concernant le statut des noms canoniques. La troisième voie échappera ou non à cette critique selon le type de théorie supplémentaire qu'elle fournira pour expliquer la nature des noms canoniques.

la signification de l'expression considérée. De ce point de vue, expliquer la validité de A / Vr(« A ») par un appel à une « faculté » de percevoir que « «$_c$ A $_c$» » dénote « A » sans autre explication n'est guère plus probant que d'expliquer la validité de A / B en s'appuyant sur une faculté à « voir » que B découle de A.[73] Ainsi, il nous semble que traiter les noms canoniques comme des primitifs dans l'analyse de la logicité des règles pour la vérité n'est pas acceptable du point de vue de la méthodologie inférentialiste.

Par sa forme, la critique des règles minimales que nous venons de développer est assez similaire à celle que Field opposait à Tarski en 1972.[74] Avant sa conversion au déflationnisme, le jeune Hartry Field avait en effet adressé une sévère objection à la théorie tarskienne de la vérité.[75] Dès les premières pages du *Wahrheitsbegriff*, Tarski se donne en effet pour objectif de construire une définition de la vérité qui ne s'appuie sur aucune notion sémantique laissée inexpliquée.[76] Ce faisant, il entendait sans

73. Nous ne voulons pas dire ici que toute explication de ce genre est par nature irrecevable. Peut-être avons-nous une « intuition » permettant de percevoir *a priori* la validité des inférences correctes ou bien une « capacité » innée permettant de saisir les lois de la logique ou certains concepts mathématiques —on pense, bien sûr, au Kant de la Raison Pure, ou bien, plus récemment, aux intuitionnistes « historiques » tels que Brouwer. Et peut-être avons-nous de même une faculté permettant de percevoir infailliblement la référence des noms canoniques. Mais, c'est exactement contre ce type d'explications que l'approche inférentialiste de la signification des constantes logiques s'est constituée... Dès lors qu'on embrasse l'approche inférentialiste de la logicité ce type d'explication devient aussitôt taboue.

Il est d'ailleurs remarquable de constater à ce sujet que, si Dummett (et dans une moindre mesure Prawitz) plaide ardemment pour l'adoption d'une logique intuitionniste en mathématiques, sa défense s'appuie sur une analyse qui, par certains aspects, est diamétralement opposée aux considérations qui donnèrent naissance à l'Ecole de Brouwer. Loin d'affirmer que la nature des objets mathématiques est d'être des constructions mentales réalisées en privé dans l'esprit du mathématicien, c'est au contraire au nom du caractère nécessairement public de la signification des énoncés mathématiques que Dummett entend identifier ce en quoi consiste la saisie de leur signification avec ce en quoi consiste la capacité à saisir ce qui en constitue une preuve. C'est à partir de ce point de départ qu'il en vient à rejeter le tiers exclu et le raisonnement par l'absurde des mathématiques classiques. On voit bien toute la différence entre les deux approches, quand bien même elles partagent les mêmes recommandations en matière de logique. Concernant le recours à des capacité primitives laissées inanalysées ou considérées comme innées, Dummett se range du côté de Frege et exclut irrémédiablement tout psychologisme dans l'analyse de la nature des lois de la logique.

74. *Cf.* Field, 1972. Bien entendu, le contexte qui nous occupe est un peu différent. Il ne s'agit pas ici de montrer que la vérité est acceptable au yeux d'un partisan du physicalisme. Ce qui nous intéresse, c'est d'étudier si la vérité peut être tenue pour une notion logique. De même, nous ne considérons pas une définition explicite de « vrai » (définition de « vrai-dans le langage objet » donnée dans un métalangage), mais une tentative de caractérisation de la signification de ce prédicat par des règles d'inférence.

75. Pour être un peu plus précis, disons que la critique de Field s'adresse non pas aux résultats mathématiques obtenus par Tarski mais plutôt à une certaine interprétation philosophique, mise en avant par Tarski lui-même, de ces résultats contenus dans Tarski, 1935.

76.

> [...] je n'ai l'intention d'utiliser pour cette construction aucun concept sémantique qui ne soit pas antérieurement réduit à quelque autre concept. (Tarski, 1935, p. 159-160)

doute donner une théorie de la vérité qui soit acceptable du point de vue de la doctrine physicaliste et réhabiliter cette notion face aux méfiances que suscitaient les concepts sémantiques chez les philosophes de son époque. Mais, nous dit Field, ce projet n'a été que partiellement couronné de succès. Le résultat positif obtenu par Tarski, à savoir la possibilité pour certains langages-objets formels $\mathcal{L}$ de définir l'expression « vrai dans $\mathcal{L}$ » dans une métathéorie, permet bien de montrer comment la valeur de vérité des énoncés logiquement complexes est déterminée par les valeurs sémantiques de leurs constituants. Toutefois, *in fine* cette construction s'appuie sur la référence des termes primitifs de $\mathcal{L}$ (ce que Field appelle la dénotation primitive) ; et cette dernière est laissée inexpliquée, elle est simplement donnée sous forme de liste. Dès lors, selon Field, Tarski n'est pas réellement parvenu à donner un fondement scientifique satisfaisant à la notion de vérité au moyen d'une réduction de celle-ci à des notions non sémantiques. Ce qu'il a réalisé se borne uniquement à réduire une notion sémantique (*i.e.* la vérité) à une autre notion sémantique (*i.e.* la dénotation primitive). Et, tant qu'on n'aura pas donné une théorie de la dénotation primitive satisfaisante du point de vue physicaliste, on ne pourra pas prétendre avoir réhabilité la notion de vérité elle-même. De façon similaire, à nos yeux, l'entreprise de réduction du prédicat de vérité à la logique, comprise ici à la lumière des canons inférentialistes, n'est parvenue qu'à des règles dans lesquelles les noms canoniques jouent un rôle crucial. Une connaissance de leurs références est un prérequis à la compréhension et à la maîtrise de ces règles. Tant qu'on n'aura pas expliqué la nature de ces noms canoniques, et montré qu'ils obéissent à leur tour à des mécanismes purement logiques, on ne peut donc pas prétendre avoir établi la logicité des règles minimales.

En résumé, dans la mesure où l'on veut montrer que les règles pour la vérité sont logiques il semble indispensable de mettre au jour explicitement tous les mécanismes dont la maîtrise est nécessaire pour saisir la signification de ce prédicat. Et, dans la mesure où le prédicat de vérité s'applique à des noms d'énoncés, fussent-ils canoniques, il apparaît donc indispensable de rendre compte du processus de nominalisation des énoncés, et d'expliquer ouvertement comment les noms canoniques sont formés et acquièrent leur référence. Il faut expliciter le mécanisme de nominalisation, en donner une théorie qui s'appuie sur ce qui est publiquement observable (du moins si on veut respecter l'esprit de l'approche inférentielle de la logicité). Postuler l'existence de noms canoniques primitifs immédiatement donnés, soit qu'on les considère comme inanalysables, soit qu'on remette leur explication à une étude ultérieure, et s'appuyer sur de tels éléments « mis de côté »

pour défendre la logicité des règles minimales ne nous semble donc pas véritablement convaincant.

2.4.2.3 Une nominalisation « quasi » logique ou syntaxique ?

Cependant, il y a peut-être une autre façon d'argumenter en faveur du déflationnisme. Plutôt que de postuler l'existence « primitive » de noms canoniques dans l'« esprit » de tous les agents cognitifs, peut-être pourrions-nous décomposer le processus de nominalistion lui-même, c'est-à-dire le passage de « A » au nom « « A » », de façon à montrer que cette opération peut elle-même être (re)construite de manière si ce n'est purement inférentielle, du moins « quasi » logique.[77] La nature « presque » logique de la vérité serait alors établie en deux étapes. La première consisterait à réduire la signification de « vrai » aux règles minimales appuyées sur des noms canoniques ; tandis que la deuxième établirait que la signification de ces derniers obéit elle aussi uniquement à des mécanismes logiques ou quasi logiques.

Bien entendu, la construction d'un nom d'énoncé n'est pas à proprement parler une inférence. Depuis Quine (1940), la notation entre guillemets permet de faire la distinction entre un énoncé en position d'usage, et un énoncé en position de citation, autrement dit mentionné.[78] Mais, l'expression « « A » » n'est pas elle-même un énoncé, une phrase déclarative susceptible de prendre la valeur Vrai ou Faux, d'être correctement ou illégitimement assertée, d'apparaître comme hypothèse ou comme conclusion dans un raisonnement, *etc.* Le passage de « A » à « « A » » ne consiste donc pas à inférer un énoncé-conclusion en partant d'un énoncé-hypothèse, mais plutôt à partir d'un énoncé pour construire une nouvelle expression, relevant d'une catégorie grammaticale distincte. On ne peut donc pas formaliser une opération de nominalisation canonique au moyen

77. Cette idée, ou quelque chose qui s'en rapproche assez, n'est pas nouvelle. On en retrouve la trace dans bien des analyses déflationnistes de la référence (voir par exemple Horwich, 1998b). Elle est aussi implicitement suggérée par Henri Galinon (2010) lorsqu'il écrit que les noms canoniques qu'il introduit sont « quelque chose comme ce qu'on obtient par une mise entre guillemets ».

78. La terminologie n'est pas fixe ici et varie selon les auteurs francophones. Il s'agit bien sûr de la fameuse distinction *use/mention* introduite par Quine *in* (Quine, 1940) avec le succès que l'on sait.

de règles structurelles données en déduction naturelle ou en calcul des séquents. [79]

Toutefois, si le procédé de nominalisation n'est pas à strictement parler une inférence, les partisans du déflationnisme peuvent tenter de se rabattre sur une position plus modeste. Pour défendre la « non-substantialité », ou la « quasi logicité », d'une notion de vérité dont les règles s'appuient sur des noms canoniques, il faut montrer que le procédé de nominalisation est le plus « innocent » possible.

En vue d'éclaircir la nature des noms canoniques, il va nous falloir en donner une théorie un peu plus explicite que ce que nous en avons dit jusqu'alors. En particulier, dès lors qu'on refuse de les prendre comme des primitifs inanalysables, une évaluation des ressources plus ou moins fortes nécessaires à l'obtention des noms canoniques ne peut faire l'économie d'une formalisation en bonne et due forme de leur construction. Mais quel est ce phénomène que, dans nos contrées, nous rendons au moyen d'une convention typographique reposant sur certains signes de ponctuation ? Au-delà de l'explication informelle : à partir de l'énoncé *xyz*, j'obtiens un nom de cet énoncé *« xyz »* par une mise entre guillemets, comment pouvons-nous rendre compte de la manière dont nous

79. On pourrait être tenté de décrire ou de caractériser l'opération de nominalisation canonique au moyen d'un schéma général semblable à ceci :

$$\text{Nominalisation/Citation} \quad -\frac{\text{A}}{\text{« A »}}- \qquad -\frac{\text{« A »}}{\text{A}}- \quad \text{Dénominalisation/Décitation}$$

Et il est alors frappant de constater que les « règles » ci-dessus se rapprochent étonnamment dans leur forme de celles utilisées en déduction naturelle pour fixer la signification des constantes logiques. Mais, il convient d'être prudent afin de ne pas faire de confusion. En réalité, le processus de nominalisation des énoncés n'est pas du tout un mécanisme purement inférentiel, semblable à ceux formalisés par les règles habituelles de la déduction naturelle. Le cadre de la déduction naturelle sert à noter les pas constitutifs de nos raisonnements déductifs. La barre horizontale (pleine) sert à noter (et à sanctionner) le passage (l'inférence) d'un énoncé (l'hypothèse) à un autre énoncé (la conclusion) —Ou peut-être de l'assertion d'une hypothèse à l'assertion d'une conclusion, ou peut-être encore d'un jugement à un autre jugement. Ici, l'expression « « A » » n'est pas elle-même un énoncé, et le passage de « A » à « « A » » n'est pas à proprement parler une inférence (d'où la ligne en pointillés).

Ce qui est bien une inférence, en revanche, ce sont les règles minimales :

$$Vr\text{-In-Min}\ \frac{A}{Vr(\text{« } A \text{ »})} \qquad \frac{Vr(\text{« } A \text{ »})}{A}\ Vr\text{-El-Min}$$

où « « *A* » » désigne le nom **canonique** de l'énoncé « *A* »

Mais ces déductions s'appuient sur une maîtrise préalable de la référence des entités « A ». Or, c'est justement ce mécanisme, rouage indispensable à l'application des règles minimales que nous voulons à présent analyser —pour percer à jour le mécanisme de l'horloge nous devons comprendre comment en fonctionnent les ressorts. C'est donc bien une explication ou une théorie de la construction de « « A » » à partir de « A » (construction qui doit permettre le passage dans les deux sens de l'un à l'autre) que nous devons fournir.

produisons les noms canoniques ? Il existe plusieurs voies pour tenter préciser cette explication informelle. La façon dont on théorisera ce processus est sans doute tributaire de la manière dont on analyse plus généralement les phénomènes linguistiques de citation ou d'autonymie. [80] Sans chercher à donner un panorama complet de toutes les conceptions philosophiques de la citation qu'on pourrait mettre au service de la nominalisation des énoncés, nous nous contenterons d'esquisser les principaux traits de celles qu'on peut qualifier de théories « syntaxiques ». [81]

Dès les années trente Tarski puis Quine [82] ont les premiers développé une théorie des guillemets intitulée théorie des citations comme noms propres. Selon cette conception, adaptée ici au cadre de notre discussion, les noms canoniques seraient des noms propres sans structure particulière désignant l'expression placée entre guillemets. Ainsi, « « la neige est blanche » » est simplement un nom propre de l'énoncé « la neige est blanche », au même titre que « Donald » est le prénom d'un célèbre canard. Du point de vue de l'analyse logique, l'expression « « la neige est blanche » » doit être considérée comme un nom simple, un mot unique qui se trouve contenir des espaces, des lettres de l'alphabet latin et des guillemets. Son comportement sémantique est similaire à celui des noms propres habituels. À ce titre, la présence de la combinaison de lettres « neige » dans l'expression « « la neige est blanche » » est aussi fortuite et sans rapport avec le fait que cette expression dénote un énoncé qui parle de la neige, que l'est la présence de la combinaison de lettres « nal » dans le prénom du personnage créé par Walt Disney, ou

80. Le recours aux guillemets de citation étant la manière standard, ou majoritairement adoptée, de noter les noms d'énoncés ou, comme on dit également souvent, les énoncés cités. Mais il existe d'autres outils que, faute de temps et d'espace, nous passerons entièrement sous silence —à l'exception des descriptions structurelles des énoncés appuyées sur une théorie de la syntaxe du langage considéré. Pour une introduction, en anglais, aux théories de la citation nous renvoyons à l'article de la Stanford Encyclopedia : Cappelen et Lepore, 2012. On pourra également consulter, en français, l'article de Philippe de Brabanter (2005). Nous nous sommes largement inspiré ici de ces deux textes.

81. et qui sont les plus favorables aux théories déflationnistes.

82. Tarski, 1935, Quine, 1940.

celle du mot « chat » dans le nom « Chatterton ».[83] Les noms canoniques sont donc des termes singuliers et il est nécessaire de connaître leurs références pour pouvoir les employer dans nos déductions. Pour savoir si Donald est marié à Daisy ou s'il a été un ardent promoteur de la guerre en Irak, il faut savoir si ce mot est employé pour désigner un canard de dessin animé ou un homme politique néo-conservateur contemporain dont le nom de famille est « Rumsfeld ». De même, pour appliquer correctement le prédicat de vérité à un nom d'énoncé, il faut savoir quel énoncé est ainsi dénommé. Mais, à vrai dire, la théorie des citations comme noms propres ne nous dit rien de particulier sur la façon dont nous pouvons fixer ou percevoir la référence des noms canoniques.

Si l'on repense à notre ordinateur de la section précédente, auquel nous voulions « enseigner la vérité », on voit que cette situation correspond au cas où nous donnerions un à un, sous forme de liste, un nom canonique pour chaque énoncé, en précisant par là leurs références. L'énoncé « la neige est blanche » reçoit le nom « « la neige est blanche » », mais ce choix est au fond arbitraire ; c'est une simple convention, et nous aurions très bien pu décider d'appeler cet énoncé « Rufus » ou « Nabuchodonosor ».

Outre qu'elle ne permet pas d'expliquer comment nous pouvons avoir accès à une infinité potentielle de noms canoniques d'énoncés, cette explication peut sembler un peu courte. Il semble clair que le nom canonique « « la neige est blanche » » a un rapport plus intime avec l'objet qu'il dénote, à savoir l'énoncé « la neige est blanche », que celui

83. Quine écrit à propos des guillemets de citation :

> Une citation n'est pas une *description*, mais un *hiéroglyphe* ; elle désigne son objet non pas en le décrivant au moyen d'autres objets, mais en le représentant. La signification du tout ne dépend pas des significations des mots qui le constituent. Le nom propre enfoui dans le premier mot de l'énoncé :
>
> « Ciceron » contient sept lettres [*« Cicero » has six letters*],
>
> par exemple, n'a sur un plan logique, rien de plus à voir avec l'énoncé que la forme verbale « tient » contenue dans le deuxième mot de l'énoncé [*is logically no more germane to the statement than is the verb « let » which is buried within the last word*]. (Quine, 1940, p. 26)

Du côté de Tarski, le passage pertinent du *Wahrheitsbegriff* est le suivant :

> Les noms obtenus par les guillemets peuvent être traités à l'instar des mots simples d'un langage, autrement dit comme des expressions syntaxiquement non composées. Les parties composantes de ces noms —-guillemets et expressions situées entre eux— remplissent la même fonction que les lettres ou les groupes de lettres se suivant les unes les autres dans les mots simples ; elles ne possèdent donc dans ce cas aucun sens autonome. Chaque expression [nom] obtenue par des guillemets est alors le nom singulier constant d'une expression déterminée (à savoir de celle qui est mise entre guillemets), nom ayant le même caractère que les noms propres des hommes. (Tarski, 1935, p. 166)

qui relie l'expression « Donald » avec telle personne ou tel personnage qui se trouvent être ainsi prénommés. Dans le cas des noms canoniques, l'objet dénoté lui-même apparaît dans l'expression qui le désigne. Les noms obtenus par une mise entre guillemets ne sont donc pas *que* des noms propres, et il y a plus dans la nominalisation canonique des énoncés que ce qui est rapporté par une liste de stipulations apparemment arbitraires. Cette incapacité à expliquer de manière entièrement satisfaisante le lien entre l'expression d'origine et sa citation est l'une des raisons pour lesquelles, malgré son importance historique, la théorie des citations comme noms propres est aujourd'hui considérée comme inadéquate et n'a plus guère de partisans [84]. À l'inverse, une théorie adéquate devra vraisemblablement prendre en compte que l'énoncé apparaît dans le nom canonique qui le désigne et que ce fait joue sans doute un rôle dans l'explication de la référence de ces derniers.

Devant de telles insuffisances, il semble donc indispensable d'en dire un peu plus sur le mécanisme qui permet de construire et d'utiliser les noms canoniques. Tarski et Quine, eux-mêmes, ont approfondi leurs explications de l'usage des noms d'énoncés et ont développé ce qui a depuis été baptisé la théorie descriptive des citations. L'idée générale est de prendre au sérieux le fait qu'une expression nommée, ou mentionnée, est présente dans le nom qui la dénote et de donner une méthode permettant de construire effectivement les noms de nos énoncés. Pour ce faire, la manière dont on va composer les noms de nos énoncés va suivre l'arbre de formation syntaxique de l'expression dénotée elle-même. On obtiendra alors des noms canoniques qui par eux-mêmes constituent autant de descriptions des expressions qu'ils sont censés désigner [85], ce qui permet d'expliquer un peu plus comment nous pouvons percevoir leur référence. Néanmoins si nous voulons donner une théorie explicite de ce processus, il nous faut introduire des ressources qui dépassent largement celles de la pure logique. Il nous faut une théorie de la syntaxe du

84. La critique la plus célèbre de la théorie des citations comme noms propres est sans doute celle de Davidson (1979) qui a depuis suscité une abondante littérature sur le sujet. Sur ce point, nous renvoyons à Cappelen et Lepore, 2012 et à Brabanter, 2005.

85. Tarksi appelle ces constructions des « structural descriptive names » (*cf.* Tarski (1935, Section 1)). Chez Quine, le passage à une théorie descriptive de la citation s'opère dans *Word and Object* :

> Au lieu de [« Tullius était romain »], nous pouvons aussi bien épeler :
>
> $T* \ u* \ l* \ l* \ i* \ u* \ s* \ espace* \ é \ * \ t* \ a* \ i* \ t* \ espace* \ r* \ o* \ m* \ a* \ i* \ n.$
>
> Nous employons alors les noms explicites de lettres et un petit signe (à la manière de Tarski) pour indiquer la concaténation. (Quine, 1999, p. 209)

Cette manière de construire des noms d'énoncés, nous dit Quine, peut être prise comme une alternative à l'usage des guillemets.

langage dont nous voulons nommer les expressions.

Plus précisément, étant donné un langage $\mathcal{L}$, on se donne $\mathcal{S}$, une théorie de la syntaxe de ce langage : elle comprendra, par exemple, un nom pour chaque symbole primitif du vocabulaire de $\mathcal{L}$[86] et un opérateur de concaténation « $*$ » pour rendre compte des combinaisons de symboles formant les expressions de $\mathcal{L}$. Pour plus de lisibilité, on dotera également $\mathcal{S}$ d'un symbole « $\sqcup$ » pour l'espace. Avec cet appareillage en main, on peut expliquer la formation des noms canoniques de la manière suivante : à supposer que $\mathcal{L}$ soit un langage écrit dans l'alphabet latin, $\mathcal{S}$ notre théorie de la syntaxe de $\mathcal{L}$ contiendra un nom pour chaque lettre latine, disons « $\overline{a}$ » pour la première de ces lettres, « $\overline{b}$ » pour la seconde, *etc.*, un symbole « $\sqcup$ » pour l'espace, et des guillemets. pour chaque symbole primitif s de $\mathcal{L}$, notre théorie $\mathcal{S}$ contient un nom de ce symbole, disons $\overline{s}$. Joint à un opérateur de concaténation, ceci permet de caractériser la classe $\mathcal{E}$ des expressions de $\mathcal{L}$, et pour chacune d'entre elles, de construire un nom qui la désigne. Les expressions de $\mathcal{L}$ sont simplement les suites de symboles obtenues par juxtaposition de symboles primitifs, et $\mathcal{S}$ nous permet de rendre compte de ce processus. Par exemple, la description structurelle suivante :

$$\overline{l} * \overline{a} * \sqcup * \overline{n} * \overline{e} * \overline{i} * \overline{g} * \overline{e} * \sqcup * \overline{e} * \overline{s} * \overline{t} * \sqcup * \overline{b} * \overline{l} * \overline{a} * \overline{n} * \overline{c} * \overline{h} * \overline{e}$$

est, dans $\mathcal{S}$, un nom de l'énoncé que nous désignons informellement par « la neige est blanche ».

En donnant explicitement le mode de construction des noms d'énoncés, la théorie descriptive des citations a le mérite de faire droit à l'idée que l'énoncé est, au moins graphiquement, présent dans son nom canonique. Ainsi, elle rend compte uniformément de la formation des noms canoniques, sans s'appuyer sur un inventaire de noms propres à la Prévert, ce qui permet, en outre, d'expliquer comment nous pouvons avoir accès à une infinité potentielle de noms canoniques (autant que l'on peut former d'énoncés dans $\mathcal{L}$).[87] Cependant ce procédé de nominalisation n'est pas un procédé logique. Une fois déroulées les ressources nécessaires pour en rendre compte, on voit que cette nominalisation requiert de maîtriser les notions syntaxiques enrégimentées dans la théorie $\mathcal{S}$. Nous sommes bien loin de règles purement structurelles données

86. Ce pourront être les lettres ou bien les mots si $\mathcal{L}$ est une langue naturelle, ou bien encore les symboles de constante, de prédicat, et les variables, parenthèses, *etc.*, si $\mathcal{L}$ est un langage formel.

87. Pour revenir de nouveau à notre exemple de la machine-apprentie, c'est sans doute ce genre de théories (au minimum) qu'il faudrait programmer dans notre ordinateur, si l'on voulait lui permettre de manipuler et de qualifier éventuellement de vrais une infinité potentielle d'énoncés.

en déduction naturelle. Si c'est bien un tel processus qui est à l'oeuvre dans l'obtention des noms canoniques, il est maintenant clair que les règles minimales ne peuvent pas sérieusement être qualifiées de logiques. Une fois dûment analysée, la mise entre guillemets ne se révèlerait alors que comme une notation commode mais dont l'apparente trivialité ne ferait que recouvrir un phénomène complexe. Derrière l'expression « « la neige est blanche » », il faudrait lire, un peu comme dans le cas d'une abréviation : « $\overline{l} * \overline{a} * \sqcup * \overline{n} * \overline{e} * \overline{i} * \overline{g} * \overline{e} * \sqcup * \overline{e} * \overline{s} * \overline{t} * \sqcup * \overline{b} * \overline{l} * \overline{a} * \overline{n} * \overline{c} * \overline{h} * \overline{e}$ ».[88]

Mais, au nom des déflationnistes, ne peut-on pas faire mieux, c'est-à-dire en l'occurrence plus simple, moins « substantiel » ? Il est vrai qu'on ne peut s'empêcher d'éprouver un sentiment de trivialité lorsqu'on regarde l'explication informelle de notre usage citationnel des guillemets. N'y a-t-il pas un moyen plus simple de rendre compte de ce mécanisme, et même, si tant est que ce soit possible, en restant plus proche de l'explication informelle ? Un examen plus minutieux de cette explication informelle pourra peut-être nous éclairer. Si l'on regarde plus attentivement ce qui semble se passer lorsque nous formons le nom d'une expression au moyen de guillemets —à partir de l'expression xyz, j'obtiens « xyz » un nom de cette expresssion par une mise entre guillemets— on s'aperçoit que cette opération possède une double nature. En réalité, deux mécanismes s'avèrent ici conjointement à l'oeuvre.

88. Les relations entre la théorie qui prend les expressions entre guillemets comme de simples noms propres et la théorie descriptive des citations ne sont pas toujours claires et varient selon les auteurs qui s'en réclament. S'agit-il de deux théories plus ou moins concurrentes cherchant à rendre compte du même phénomène ? La théorie des descriptions serait alors une amélioration, ou une variante plus poussée, de la théorie des noms propres. Mais les deux théories voudraient en fait expliquer le même phénomène, et la notation typographique utilisant les guillemets ne préjugerait pas de la nature des noms d'énoncés. Ou bien faut-il faire la différence entre deux procédés distincts de nominalisation : les descriptions structurelles et les guillemets de citation ? Dans la première section du *Wahrheitsbegriff*, Tarski introduit deux théories distinctes quoique reliées. Mais il ne donne pas de théorie plus précise des « quotation-mark names » (c'est un foncteur...). Chez Quine, du moins dans le passage de Quine, 1999 que nous avons cité, les deux théories sont présentées comme des alternatives dont les usages sont équivalents. Pour Geach en revanche, (Geach, 1957), l'un des premiers à avoir vertement critiqué les insuffisance de la théorie des noms propres, la théorie descriptive est présentée comme une élaboration et amélioration de la théorie des noms propres. Donc, pour lui, l'expression « la neige est blanche » est bien ultimement analysée —et doit être comprise— comme $\overline{l} * \overline{a} * \sqcup * \overline{n} * \overline{e} * \overline{i} * \overline{g} * \overline{e} * \sqcup * \overline{e} * \overline{s} * \overline{t} * \sqcup * \overline{b} * \overline{l} * \overline{a} * \overline{n} * \overline{c} * \overline{h} * \overline{e}$.

Sur ce point, voici ce que dit Brabanter (2005, p. 10) :

> [...] Tarski et Quine ont, parallèlement à la doctrine de l'autonyme comme nom, développé une théorie extensionnellement équivalente — que certains auteurs postérieurs nommeront « Théorie descriptive » de l'autonymie — qui voit dans l'autonyme l'abréviation d'une description telle que *Le cinquième mot du poème « The Raven »* ou encore d'une description « orthographique » comme *Le mot composé de 'b','o','s','t','o','n', dans cet ordre*. En plusieurs endroits Quine et Tarski qualifient ces descriptions de *noms* [...] et il semble bien qu'ils fassent un usage interchangeable des termes de *nom* et *description*.

1. *Processus syntaxique.* Un premier processus permet de construire une nouvelle expression du langage considéré. Ceci s'apparente à une simple manipulation de signes. À partir d'une suite de symboles « xyz », je construit une nouvelle suite de symboles de la manière suivante, un guillemet ouvrant suivi de ma suite de symboles suivi d'un guillemet fermant, avec pour résultat l'expression « « xyz » ». Schématiquement :

$$\begin{array}{ccc} xyz & \mapsto & \text{« } xyz \text{ »} \\ \exists xRxy & \mapsto & \text{« } \exists xRxy \text{ »} \\ \text{la neige} & \mapsto & \text{« la neige »} \end{array}$$

Ce premier processus est incontestablement un mécanisme syntaxique. Les expressions guillemetées sont obtenues par concaténation de symboles, au même titre que je forme l'énoncé « $\exists xRx$ » à partir de la formule ouverte « Rx ». Jusque là, rien d'extraordinaire. Mais l'explication informelle de la mise entre guillemets ne se réduit pas à cela. En stipulant que ma nouvelle expression est un nom de mon expression de départ, je précise également son comportement sémantique.

2. *Processus sémantique.* Le second mécanisme à l'oeuvre est donc celui qui permet de fixer la référence de l'expression nouvellement formée : « « xyz » » dénote « xyz », « « $\exists xRxy$ » » dénote « $\exists xRxy$ », « « la neige » » dénote « la neige », *etc.* En toute généralité, ceci revient à stipuler qu'une expression formée par un guillemet ouvrant suivi d'une suite de symboles, notons la s, suivie d'un guillemet fermant dénote s.[89]

À partir de cette quasi paraphrase de l'explication informelle, dans laquelle on a néanmoins pris soin de distinguer les deux mécanismes en jeu, certains auteurs ont tenté

89. Ce type d'analyses de la mise entre guillemets qui distingue minutieusement entre une part proprement syntaxique et une part sémantique est due à Mark Richard Richard, 1986, p. 390 [nous traduisons] :

> [...] considérant comment on constuirait une théorie tarskienne de la vérité pour des langages contenant des « citations à la Quine » et des « citations à la Tarksi ». Chaque théorie contiendrait des axiomes énoncés à peu près comme ceci :
>
> (A) Pour toute suite de symboles e, un guillemet ouvrant suivi de e suivi d'un guillemet fermant est un terme individuel.
>
> (B) Pour toute suite de symboles e, un guillemet ouvrant suivi de e suivi d'un guillemet fermant dénote e.

La version fonctionnelle de ces deux processus est une élaboration due à Ludwig et Ray (1998, Note 43, p. 163). Ces deux sources, sont citées dans Cappelen et Lepore, 2012, où elles sont exposées à titre de représentants des théories décitationnelles de la citation (« Disquotational Theory of Quotation »).

d'élaborer une formalisation en bonne et due forme de la mise entre guillemets. D'après une suggestion de Ludwig et Ray, 1998, au double processus ci-dessus correspondent deux fonctions. Une première fonction syntaxique, notons la « $\mathcal{N}om_{«»}$ », expose le mode de formation des noms de nos expresssions, tandis qu'une seconde, notée « $\mathcal{R}ef_{«»}$ », permet de contrôler le comportement sémantique d'une classe d'expressions particulières, à savoir les expressions guillemetées. Selon cette analyse, les noms canoniques ne sont donc ni des noms propres sans lien particulier avec l'expression qu'ils dénotent, ni des descriptions structurelles. Il faut plutôt voir la mise entre guillemets comme une double opération fonctionnelle sur l'ensemble des expressions du langage, chaque opération prenant comme argument une expression et donnant en sortie une autre expression. Mais que se passe-t-il si, en vue d'évaluer précisément les ressources exigées par la nominalisation canonique, nous voulons une théorie entièrement formalisée de ces deux fonctions ? Pour définir ces dernières, il faut que nous puisssions caractériser leur domaine et identifier les objets ou les termes auxquels elles s'appliquent. En l'occurrence cela revient à caractériser la classe des expressions du langage considéré et à être capable d'identifier chacune d'entre elles. Autrement dit, pour formaliser la mise entre guillemets telle qu'elle est comprise par Richard, 1986 et par Ludwig et Ray, 1998, nous aurons besoin de quelque chose comme une théorie syntaxique du langage dont nous cherchons à nommer (canoniquement) les énoncés.

Plus explicitement, voici à quoi pourrait ressembler une axiomatisation poussée à son terme des fonctions évoquées par Ludwig et Ray, 1998 [90] : soit $\mathcal{E}$ l'ensemble des expressions de notre langage ; si ce n'est déjà fait, on se munit également d'un opérateur de concaténation $*$ sur les expressions dans $\mathcal{E}$, et on désigne nos guillemets par les symboles $«_c$, et $_c»$, les deux fonctions ci-dessus sont alors définies par les axiomes suivants :

$$1.\ \forall e \in \mathcal{E},\ \mathcal{N}om_{«»}(e) = «_c\ *\ e\ *\ {}_c»$$
$$2.\ \forall e \in \mathcal{E},\ \mathcal{R}ef_{«»}(«_c\ *\ e\ *\ {}_c») = e$$

Ainsi, même la version fonctionnelle de la mise entre guillemets requiert de maîtriser

90. Notez que le même type de ressources s'imposeraient si nous voulions formaliser complètement les axiomes (A) et (B) de Richard, 1986 (*cf.* la note précédente, n° 89). Même s'il formule ses axiomes en langue naturelle, lorsque Mark Richard quantifie sur les suites de symboles e et leur juxtapose des guillemets ouvrant et fermant, cela signifie qu'il dispose implicitement d'un moyen d'identifier ces objets e et qu'il maîtrise une opération de concaténation sur ces objets .

suffisamment la syntaxe de notre langage pour pouvoir en identifier les énoncés, considérés comme de simples suites de symboles, et pouvoir manipuler ces suites de façon à y accoler des symboles de guillemets. L'usage implicite d'une théorie de la syntaxe du langage considéré transparaît ici à travers la quantification sur $\mathcal{E}$, et la construction des termes du type «$_c$ $*$ e $*$ $_c$» . Si nous voulions donner un sens plus précis à $\mathcal{E}$ et expliquer comment ses éléments sont construits, il nous faudrait avoir recours à une théorie syntaxique sans doute similaire à celle que nous avons introduite lors de notre exposition de la théorie des descriptions structurelles. [91]

Quelle leçon retenir de tout ceci pour le déflationnisme ? Le très rapide tableau que nous venons de brosser de ce que pourrait être une formalisation des processus qui sous-tendent l'usage des noms canoniques n'a pas la prétention d'être exhaustif. Nous nous sommes sciemment cantonnés aux théories des guillemets les plus favorables aux thèses déflationnistes. [92] Néanmoins, quelle que soit la manière dont, au final, on théorisera l'opération de nominalisation canonique de nos énoncés, il semble qu'on ne puisse pas faire l'économie d'une théorie de la syntaxe de notre langage.

Dans la section précédente, nous avons expliqué pourquoi, dans l'entreprise d'évaluation de l'éventuelle logicité d'un prédicat de vérité déflationniste, s'appuyer sur des noms

91. Henri Galinon évoque ce problème :

> Il y a, me semble-t-il, un sens dans lequel chaque énoncé du langage possède un nom canonique véritablement transparent : quelque chose comme ce que l'on obtient par la mise entre guillemets de cet énoncé. Bien entendu, une analyse possible de ce qu'est, en fait, la mise entre guillemets, nous ramènerait tout droit aux descriptions structurelles : l'énoncé entre guillemets est simplement un nom de l'énoncé obtenu en utilisant les symboles comme noms d'eux-mêmes et la juxtaposition comme opérateur de concaténation. (Galinon, 2010, p. 317)

il poursuit ensuite :

> Mais cette reconstruction rationnelle ne rend pas compte du caractère d'immédiateté du processus par lequel je suis capable de former des noms d'énoncés de mon langage par la mise entre guillemets, et de reconnaître un énoncé sous un nom de ce genre. (Galinon, 2010, p. 317)

avant de proposer de considérer les noms canoniques comme des primitifs. Mais, sur ce dernier point, nous avons déjà fait part de nos réserves quant à ce qui nous semble n'être qu'une échappatoire.

92. Les plus favorables parce que ce sont les théories les plus « modestes », dans la mesure où elles se proposent d'analyser la nominalisation canonique à partir de ressources relativement faibles (mais qui dépassent néanmoins le cadre de la logique). Il existe d'autres conceptions de la citation qui sont bien moins favorables, pour ne pas dire entièrement défavorables, à l'idée que l'opération de mise entre guillemets puisse être rapprochée de quelque manière que ce soit d'une opération logique ou quasi logique, ou même qu'on puisse la considérer comme une opération relevant seulement de la syntaxe. Nous pensons par exemple à la théorie démonstrative de la citation, chère à Davidson, ou à la théorie de la citation identité. Nous nous permettons de renvoyer de nouveau le lecteur qui voudrait s'en convaincre à Cappelen et Lepore, 2012 et à Brabanter, 2005.

canoniques considérés comme primitifs ne nous semblait pas acceptable. Après examen des mécanismes gouvernant la nominalisation de nos énoncés, on constate que, malgré le passage par des règles minimales, on est quand même conduit à ne *pas* considérer le prédicat de vérité comme une notion purement logique, puisque son emploi réclame de maîtriser, à tout le moins, une théorie syntaxique. Sa signification n'est pas réductible entièrement à des règles purement structurelles en harmonie. L'hypothèse des noms canoniques primitifs ne faisait que masquer l'importance des ressources nécessaires à leur obtention. Parallèlement, on peut aussi remarquer que ces ressources indispensables à la nominalisation, même canonique, nous donnent des extensions non conservatives sur la logique pure. [93] Les règles ne satisfont donc pas non plus l'harmonie globale. Ainsi, tout bien considéré l'élagage des règles générales qui mène aux règles minimales ne permet pas de réellement résoudre la double difficulté que posent les règles pour « vrai » dans la tentative de reconcilier la vérité avec le critère inférentialiste de logicité.

Néanmoins, pour la question qui nous occupe, il est un point intéressant propre aux théories syntaxiques de la nominalisation canonique que nous venons d'examiner. Si on laisse de côté la théorie des noms propres, qui ne nous dit rien sur la manière dont sont construits les noms d'énoncés, pas plus que sur le lien rattachant un nom canonique à l'énoncé qu'il désigne, on constate que dans la théorie des descriptions structurelles, tout comme dans l'analyse fonctionnelle de la mise entre guillemets, la construction des noms se fait indépendamment du contenu sémantique de l'énoncé. L'énoncé est identifié de manière purement graphique. Des notions aussi douteuses que le sens de l'énoncé, son contenu sémantique ou propositionnel n'entrent pas en jeu dans la construction du nom canonique de cet énoncé. Bien sûr, il faut bien admettre que la nominalisation n'est pas un procédé logique, purement inférentiel qu'on pourrait formaliser par des règles données en déduction naturelle ou dans le cadre d'un calcul des séquents. C'est, de fait, un phénomène plus complexe, qui réclame une théorie syntaxique. Mais, si on s'en tient aux formalisations que nous avons évoquées, cela reste un mécanisme purement calculatoire, une simple manipulation de symboles contrôlée par des règles, qu'on pourrait qualifier de formelles, même si elles sont d'un niveau plus complexe que les règles gouvernant

93. Nous avons déjà évoqué ce résultat. Mais à présent nous voyons plus clairement pourquoi l'introduction des noms canoniques nous fait sortir du périmètre de la logique. Nous savions qu'elle s'accompagne d'engagements ontologiques concernant l'existence de nouvelles entités (les noms eux-mêmes) ; mais si nous refusons de simplement les prendre comme primitifs, nous voyons aussi quel genre de théories leur explication réclame. Ces théories sont clairement hors de l'espace des simples règles, ou lois, de la logique, *a fortiori* si on adopte un critère inférentialiste de logicité.

les constantes logiques. Soulignons que, sur ce point, la théorie fonctionnelle de la mise entre guillemets est particulièrement radicale et efficace. En effet, selon cette analyse, non seulement la construction d'un nouveau terme opérée par la fonction $\mathcal{N}om_{«»}$ est évidemment de nature syntaxique, mais la fonction $\mathcal{R}ef_{«»}$ donnant la référence des noms canoniques est en un sens elle aussi réduite à une opération syntaxique : la référence du nom «$_c$ $*$ e $*$ $_c$», c'est-à-dire l'expression e, *qua* combinaisons de symboles, peut être obtenue par une simple suppression des guillemets extérieurs. Une connaissance de la référence peut donc se ramener à une capacité de manipuler des symboles, indépendamment de leur signification.[94] Après tout, n'est-ce pas ce que nous faisons lorsqu'au cours de notre lecture, notre regard se pose sur une expression guillemetée ? Nous voyons aussitôt quelle est sa référence en effaçant mentalement les guillemets qui l'entourent.

Au vu du rôle crucial joué par la syntaxe dans le processus de nominalisation, la ligne d'argumentation suivante a donc été proposée au nom des déflationnistes : le mécanisme de nominalisation n'est pas à proprement parler une opération logique, mais ce serait une opération purement... syntaxique. On obtiendrait les noms de nos énoncés par une simple manipulation de symboles et de signes. Dès lors, un prédicat de vérité déflationniste, dont la signification est donnée par des règles minimales articulées sur un processus de nominalisation de cette nature, hériterait de ces qualités de modestie. La vérité, simple outil expressif de décitation, n'est pas à strictement parler une constante logique, mais ce serait une notion logico-syntaxique.[95] C'est là une thèse amendée, plus faible, que celle qui consiste à prendre la vérité pour une notion logique. Mais cette thèse remaniée reste fidèle à l'idée déflationniste selon laquelle la vérité ne serait pas une propriété « substantielle »[96], sans pour autant tenter à toute force de contraindre

94. En passant, remarquons que la composition de ces deux fonctions donne, dans un sens, l'identité sur $\mathcal{E}$:

$\forall e \in \mathcal{E},\ (\mathcal{R}ef_{«»} \circ \mathcal{N}om_{«»})(e) = \mathcal{R}ef_{«»}(«_c\ *\ e\ *\ _c») = e$

dans l'autre l'identité sur l'ensemble des expression guillemetées :

$\forall e \in \mathcal{E},\ (\mathcal{N}om_{«»} \circ \mathcal{R}ef_{«»})(«_c\ *\ e\ *\ _c») = \mathcal{N}om_{«»}(e) = «_c\ *\ e\ *\ _c»$

95. On trouve cette appellation chez certains auteurs, notamment logiciens, cherchant à caractériser plus rigoureusement la thèse déflationniste selon laquelle le prédicat de vérité est un outil de décitation, privé de tout contenu « substantiel », sans pour autant l'identifier à une notion logique (*cf.* par exemple Halbach (2001b) et Horsten (2011)). On trouve aussi parfois l'appellation d'outil logico-mathématique ou logico-arithmétique puisque, bien souvent dans les approches formelles du prédicat de vérité, c'est une théorie arithmétique où les énoncés sont désignés par leurs numéros de Gödel qui tient lieu de théorie syntaxique.

96. Étant donné le caractère flou de cette appellation, une certaine latitude est sans doute permise dans la manière dont on la comprend. Devant l'échec, dans le cas de la vérité déflationniste, de l'identification stricte de la non substantialité à la logicité, on peut donc se rabattre sur une autre « traduction » de l'absence de substantialité : non-substantiel = purement syntaxique.

le prédicat « vrai » à remplir le critère inférentiel de logicité.

Toutefois, l'histoire ne s'arrête peut-être pas là. Même cette thèse plus faible, selon laquelle l'opération de nominalisation/citation des énoncés serait une opération purement syntaxique, ce qui ferait du prédicat de vérité un outil logico-syntaxique, est sujette à caution.

2.4.2.4 Quel contenu pour les noms canoniques ?

Dans un article paru en 2004,[97] François Rivenc soulève en effet un problème pour ce qu'il appelle le déflationnisme linguistique et propose une critique de ce qu'il intitule l'*interprétation purement citationnelle des guillemets*. Nous voudrions reprendre ici une partie de ses observations qui nous semblent pertinentes pour notre examen des noms canoniques. Le coeur de l'argument est le suivant : pour que l'analyse déflationniste de la notion de vérité puisse être correcte, pour que les règles minimales soient intuitivement valides, il est indispensable que les noms canoniques ne soient pas *que* des constructions purement syntaxiques.[98]

Considérons, une fois encore, une instance des règles minimales :

$$Vr\text{-In-Min}\ \frac{\text{la neige est blanche}}{\text{«}_c\text{ la neige est blanche }_c\text{» est vrai}}\qquad \frac{\text{«}_c\text{ la neige est blanche }_c\text{» est vrai}}{\text{la neige est blanche}}\ Vr\text{-El-Min}$$

97. Rivenc, 2004.

98. L'argument se trouve précisément aux pages 520-521 de (Rivenc, 2004). Il porte sur une version plus habituelle de la théorie déflationniste de la vérité, à savoir l'axiomatisation de cette notion au moyen d'une collection infinie de T-équivalences (précisément les T-équivalences décitationnelles de la forme : « p » est vrai ssi p). Le raisonnement de François Rivenc s'attaque à la question de la nature analytique (*i.e.* vraie en vertu du seul sens des mots) des T-équivalences et s'appuie sur une comparaison entre les T-équivalences homophoniques et hétérophoniques, que nous ne reprenons pas ici. Notons également que Rivenc ne précise pas ce qu'il entend au juste par interprétation citationnelle des guillemets, même s'il semble clair qu'il a en tête les propos de Quine à ce sujet, c'est-à-dire, très vraisemblablement, la théorie des citations comme noms propres. Nous avons donc un peu modifié son argument pour l'adapter à notre cadre d'analyse et à la discussion précise qui nous intéresse ici. Mais l'idée centrale reste la même : le contenu sémantique d'un énoncé doit être « actif » dans le nom (pour ce qui nous concerne : dans le nom canonique) de cet énoncé ; ce dernier ne peut donc pas uniquement désigner une simple suite de symboles, c'est-à-dire dénoter l'expression citée seulement en tant que suite de symboles.

Voici les mots de conclusion de François Rivenc :

> L'analycité des équivalences T (quand elles sont analytiques !) provient du fait que les énoncés ne sont pas simplement cités, mais utilisés, quoique de manière déviante. (Rivenc, 2004, p. 521)

Cette paire de règles est censée fixer la signification du prédicat de vérité. Elles sont donc valides, si l'on veut, par définition. Nous pourrions ajouter, de surcroît, qu'elles nous paraissent également intuitivement correctes, ce qui fonde la plausibilité initiale du déflationnisme.[99] Mais, ce que ces règles saisissent, la relation qu'elles entérinent entre les deux énoncés concernés, est ce qu'on appelle une relation d'équiassertabilité : la licence qui nous est donnée de passer d'une ligne à l'autre traduit le fait que toute justification de l'un vaut justification de l'autre. Quel que soit le « chemin » par lequel nous sommes parvenus à « la neige est blanche » :

$$\begin{array}{c}\vdots\\ \text{la neige est blanche}\end{array}$$

la règle d'introduction pour « vrai » nous garantit que ce même « chemin » nous mène à « «$_c$ la neige est blanche $_c$» est vrai » :

$$Vr\text{-In-Min}\ \frac{\begin{array}{c}\vdots\\ \text{la neige est blanche}\end{array}}{\text{«}_c\text{ la neige est blanche }_c\text{» est vrai}}$$

Il en est de même, *mutatis mutandis*, pour la règle d'élimination.

Or, et c'est là la deuxième étape du raisonnement, il semble clair que la justification d'un énoncé donné va dépendre du contenu sémantique particulier de cet énoncé.[100] Sans doute, ce qui constitue une justification de l'énoncé « la neige est blanche » doit avoir affaire avec la blancheur de la neige, ce qui constitue une justification de l'énoncé « l'herbe est verte » doit avoir affaire avec la couleur de l'herbe, *etc.* De par l'équiassertabilité, ceci a pour conséquence que toute justification de l'énoncé « «$_c$ la neige est blanche $_c$» est vrai » va dépendre du contenu sémantique de l'énoncé « la neige est blanche », toute justification de l'énoncé « «$_c$ l'herbe est verte $_c$» est vrai » va dépendre du contenu sémantique de l'énoncé « l'herbe est verte », *etc.* Tout cela est d'ailleurs en parfait accord avec certaines analyses déflationnistes. Ainsi, lorsque Quine nous dit que le prédicat de vérité « annule l'effet des guillemets et nous ramène vers le monde », lorsqu'il nous dit qu' « en attribuant la vérité à l'énoncé «$_c$ la neige est blanche $_c$» j'attribue la blancheur

99. La question étant alors bien sûr de savoir si ces règles valides sont constitutives de la signification de la vérité, et si elles en offrent une analyse exhaustive, ou s'il y a d'autres choses à dire sur la vérité dont ces règles elles-mêmes seraient peut-être dérivées.

100. Nous avons usé librement de l'expression « contenu sémantique » (on aurait pu dire « contenu propositionnel ») d'un énoncé, sans préciser le sens de cette expression.

à la neige » [101], cela semble bien suggérer que le contenu sémantique des deux énoncés est, si ce n'est rigoureusement identique, du moins grandement similaire [102] et que toute justification d'un des deux énoncés vaudra justification de l'autre.

Le problème c'est que ceci n'est pas sans conséquence sur la nature des noms canoniques. Pour que ce qui constitue une justification de « «$_c$ A $_c$» est vrai » dépende du contenu sémantique de « A », il faut bien que ce contenu soit présent quelque part dans l'expression « «$_c$ A $_c$» est vrai ». Or, où peut-il se cacher ? Sans doute pas dans la locution « est vrai ». La seule chose qui varie entre, par exemple, les énoncés « «$_c$ la neige est blanche$_c$» est vrai » et « «$_c$ l'herbe est verte $_c$» est vrai » sont les noms canoniques qui y figurent. C'est donc bien dans ces noms que le contenu sémantique de l'énoncé nominalisé doit se loger.

Autrement dit, ce qui est visé par un nom canonique n'est pas seulement un énoncé, en tant que combinaison de symboles donnée par une théorie syntaxique du langage considéré. Au contraire, le contenu sémantique d'un énoncé est toujours « actif » dans le nom canonique qui le désigne. Ceci conduit François Rivenc à dire que l'énoncé graphiquement présent dans le nom formé au moyen des guillemets [103] n'est, en fait, pas simplement cité mais bel et bien utilisé, quoique de manière déviante. Or, on ne voit pas comment une telle chose serait possible si la nominalisation canonique se résumait à une pure manipulation syntaxique de symboles, qui s'effectuerait indépendamment du sens des mots et des énoncés que ces derniers servent à construire.

Pour bien voir où se situe la difficulté, comparons avec une situation dans laquelle, *ex hypothesi*, les guillemets ont un usage purement citationnel, au sens où nous sommes dans un cas où l'expression citée est bien considérée uniquement comme une suite de symboles. [104] Depuis que Quine a introduit la distinction usage/mention, il est de coutume de présenter cette démarcation aux étudiants dès les premiers cours de logique philoso-

101. Quine (1990, p. 116-117), voyez le chapitre précédent pour plus de détails.

102. Certaines analyses déflationnistes sont encore plus radicales et considèrent que « A » et « «$_c$ A $_c$» est vrai » ne sont que des variantes notationnelles, deux façons différentes de noter le même énoncé. Le contenu de ces deux expressions est alors considéré comme une seule et même chose.

103. Pour ce qui nous concerne, disons : présent dans son nom canonique.

104. Attention ! Remarque importante : nous avons déjà exposé dans la section précédente certains problèmes que l'on rencontre lorsqu'on essaye de rendre compte des phénomènes linguistiques de citation ou d'autonymie. En particulier, nous avons rappelé les insuffisances de certaines analyses historiques de l'usage des guillemets (notamment celle(s) de Quine). Ici, la démarche est un peu différente. Sans préjuger de ce qu'est en réalité une bonne théorie des guillemets, ou une bonne théorie de la nominalisation des énoncés, nous voudrions voir si une conception purement syntaxique de ces phénomènes peut convenir dans l'analyse de la signification du prédicat de vérité. Par hypothèse, nous supposons donc ici qu'une telle théorie est possible, et qu'elle est au moins à première vue adéquate.

phique. On utilise alors les guillemets pour signaler qu'une expression est mentionnée et non pas employée. En d'autres termes, on signale par des guillemets les usages autonymes d'une expression. Et l'on insiste bien auprès des étudiants pour qu'ils ne fassent pas de confusion. Ainsi, les énoncés :

« Aristote » contient huit lettres.

Aristote fut le précepteur d'Alexandre le Grand.

sont corrects. Mais si nous écrivons : « Aristote » a écrit les *Seconds Analytiques*, nous commettons une erreur grave. La personne physique du philosophe antique n'est pas présente dans la suite de lettres que l'expression « « Aristote » » désigne. De même, il serait absurde d'affirmer qu'Aristote, fait de chair et d'os, contient la lettre « A ». Tout ceci est élémentaire et bien connu.[105] Le point crucial est que la valeur sémantique de l'expression utilisée, en l'occurrence un certain philosophe antique, n'a rien à voir avec celle de l'expression mentionnée, *i.e.* une certaine combinaison de huit lettres qui se trouve entourée de guillemets. Le contenu sémantique de l'expression « Aristote » *s'est complètement perdu* lors du passage à l'expression « « Aristote » ». Pour reprendre la terminologie de Rivenc, dans une interprétation purement citationnelle des guillemets, « « Aristote » » désigne une simple suite de lettres, tout comme « « pfffrg » » désigne « pfffrg ». Or, il nous semble que c'est bien ce type d'interprétation qui est formalisé par ce que nous avons appelé les théories syntaxiques de la citation. Chez Quine, ce point est assez clair : les expressions guillemetées dénotent simplement les suites de symboles mentionnées. Mais, de la même manière, les descriptions structurelles sont totalement silencieuses sur la signification des expressions qu'elles désignent. Enfin, dans l'interprétation fonctionnelle de la mise entre guillemets, la fonction $\mathcal{R}ef_{\text{«»}}$ nous donne bien comme référence de l'expresssion « «$_c$ e $_c$» » la suite de symboles e. Mais elle ne nous dit rien sur le sens ou le contenu sémantique de e lui-même. Si c'est bien un phénomène de cette nature qui est à l'oeuvre dans la nominalisation canonique de nos énoncés, on ne voit pas comment le contenu sémantique de « la neige est blanche » pourrait jouer le moindre rôle dans les justifications possibles de l'énoncé « «$_c$ la neige est blanche $_c$» est vrai ».

Comparons avec $A \wedge B$ ici A et B apparaissent explicitement. Dans, « «$_c$ A $_c$»est vrai », ce qui apparaît n'est pas l'énoncé « A » lui-même, mais un nom de cet énoncé.

105. Même si, répétons-le, nous ne préjugeons pas dans cette section de ce que serait la bonne manière d'en rendre compte.

Qu'on puisse néanmoins retrouver le contenu de « A » à partir du seul énoncé « «$_c$ A $_c$»est vrai », montre que la référence de « «$_c$ A $_c$» » n'est pas indépendante du contenu de « A ».

Dans le même ordre d'idée, si nous réexaminons l'analyse de l'inférence :

$$\frac{\text{La dernière phrase prononcée par Napoléon lors de son allocution à Embabèh est vraie}}{\text{Du haut de ces pyramides quarante siècles vous contemplent}}\ Vr\text{-El-générale}$$

où, à la suite d'Henri Galinon, nous isolions le rôle joué par les règles minimales [106] :

$$\text{PSI}\ \frac{\dfrac{\text{Vr}(\beta) \qquad \beta = \text{«}_c \text{ Du haut de ces pyramides quarante siècles vous contemplent } _c\text{»}}{Vr(\text{ «}_c \text{ Du haut de ces pyramides quarante siècles vous contemplent } _c\text{»})}}{\text{Du haut de ces pyramides quarante siècles vous contemplent}}\ Vr\text{-El-Min}$$

on peut se demander si l'identification $\beta =$ «$_c$ Du haut de ces pyramides ...$_c$», n'est bien qu' une identification « syntaxique ». Car ce n'est pas une égalité entre combinaison de signes qui justifie cette inférence, c'est bien une identité entre contenus sémantiques. [107] Lorsque nous posons l'égalité entre la référence de la description définie « la dernière ... Embabèh » et celle du nom canonique « «$_c$ Du haut ... contemplent $_c$» » [108], nous ne voulons pas dire, ou pas seulement dire, que Napoléon a prononcé une certaine combinaison syntaxique de phonèmes, indépendamment du sens de ce qu'il a dit. Ce que nous voulons identifier, c'est le contenu sémantique de l'énoncé prononcé il y a plus de deux siècles par Napoléon avec celui de notre propre énoncé, dénoté par le nom canonique que nous transcrivons à l'écrit sur cette page. D'ailleurs, que se passe-t-il si Napoléon avait un défaut de prononciation ? S'il bégayait ? S'il avait l'accent corse ? S'il était sujet aux fautes d'orthographe ? Plus radicalement, supposons que Napoléon ait parlé non pas le français, mais un langage similaire en tout point au français, à ceci près que dans ce langage le mot « siècle », pris à la fois comme combinaison de phonèmes et comme une

106. On note β une abréviation de « La dernière phrase prononcée par Napoléon lors de son allocution à Embabèh »

107. Au demeurant, on peut se demander en quoi consiste exactement une identification « syntaxique ». S'agit-il d'une identité entre combinaisons de signes prises comme des *tokens* (auquel cas l'égalité présente dans la première ligne de l'inférence ci-dessus est trivialement fausse, puisque le *token* prononcé par Napoléon n'est évidemment pas le même que celui couché sur cette page) ? S'agit d'une égalité de *types* ? Mais alors comment ces derniers sont-ils définis : comme combinaison de phonèmes, de graphèmes, de lettres ... ?

108. À ce titre, la notation $\beta =$ «$_c$ Du haut de ces pyramides quarante siècles vous contemplent $_c$» est un peu abusive puisque ce ne sont pas les noms eux-mêmes qui sont (censés être) égaux, mais leur(s) référence(s).

certaine suite de lettres, désigne non pas une période de temps mais une certaine espèce de sauterelles très commune dans le désert égyptien et dont quarante membres se trouvaient au sommet des pyramides, les yeux tournés vers les soldats de la (future) Grande Armée. Dans ce scénario fictif, l'inférence ci-dessus devient évidemment incorrecte. Et c'est bien l'identification :

$$\beta = \text{«}_c \text{ Du haut de ces pyramides quarante siècles vous contemplent } _c\text{»}$$

qui pose problème, puisque notre autre hypothèse de départ, $\mathrm{Vr}(\beta)$, est vraie —à supposer que quarante sauterelles étaient bien présentes sur les pyramides et qu'elles contemplaient les soldats— et que les inférences suivantes ne sont que des applications de règles que l'on supposera adéquates puisqu'il s'agit du principe de substitution des identiques (PSI) et d'une règle d'Elimination minimale. Or, pourtant, dans notre scénario, Napoléon a bien prononcé la bonne combinaison syntaxique de phonèmes. Mais bien sûr, il n'a pas voulu dire par cet acte ce que, nous, nous disons en prononçant ou en écrivant les mots : « du haut de ces pyramides quarante siècles vous contemplent ». En somme, la dernière phrase prononcée par Napoléon dans notre scénario fictif n'est pas la même que celle de notre français standard bien qu'elle soit composée de la même combinaison de signes, c'est-à-dire en l'occurrence de la même combinaison de phonèmes. [109] Et, l'on voit bien que la seule identité de forme syntaxique ne suffit pas pour caractériser adéquatement la référence des noms canoniques. Ce que ces derniers désignent, ce n'est pas un énoncé, *qua* combinaison syntaxique, c'est bien plutôt son contenu sémantique.

Il existe une réponse assez naturelle dont le déflationniste peut se prévaloir. Dans notre scénario fictif, Napoléon ne parle pas le même langage que celui que nous utilisons. Nous avons donc changé de langage au cours de notre inférence. Il n'est dès lors pas étonnant qu'elle déraille ! Pour éviter ce genre de désagrément, les déflationnistes prennent la précaution de préciser qu'un locuteur ne doit employer le prédicat de vérité qu'avec des (noms d') énoncés de son propre idiolecte, c'est-à-dire qu'il ne doit attribuer la vérité qu'à des énoncés qu'il comprend et qui font partie de son propre langage. [110] Les noms canoniques ne s'emploient par conséquent que pour désigner des énoncés de notre propre idiolecte. Ce n'est que dans un second temps que l'on peut, ensuite, généraliser

109. Bien évidemment, nous pourrions tout aussi bien imaginer un scénario dans lequel Napoléon aurait écrit la bonne combinaison syntaxique de symboles écrits, tout en lui attribuant un sens déviant de celui du français. Mais nous avons voulu conserver notre exemple.

110. Ceci est explicitement dit dans (Horwich, 1998b) tout comme dans (Field, 1994b) et (Field, 1994a).

les usages de « vrai » en recourant à des traductions. Dans le cas de notre Napoléon imaginaire et de son langage déviant, la bonne analyse de l'inférence :

$$\frac{\text{La dernière phrase prononcée par Napoléon lors de son allocution à Embabèh est vraie}}{\text{Du haut de ces pyramides quarante siècles vous contemplent}}\ Vr\text{-El-générale}$$

devrait donc prendre en compte le caractère « non-idiolectique » de la langue de Napoléon. Il faudrait donc prendre garde à bien traduire préalablement l'énoncé prononcé par Napoléon dans notre propre langage. Ainsi, une analyse correcte pourrait, par exemple, avoir la forme suivante :

$$\text{PSI}\ \frac{\text{Vr}(\beta) \qquad \beta \equiv \text{«}_c\ \text{Du haut de ces pyramides quarante sauterelles vous contemplent}\ _c\text{»} \pmod{trad.}}{\dfrac{Vr(\ \text{«}_c\ \text{Du haut de ces pyramides quarante sauterelles vous contemplent}\ _c\text{»})}{\text{Du haut de ces pyramides quarante sauterelles vous contemplent}}\ Vr\text{-El-Min}}$$

où (mod *trad.*) signale que l'on a pris soin préalablement de traduire l'énoncé de Napoléon dans notre propre idiolecte. [111]

Mais cette réponse n'est pas vraiment satisfaisante au regard du problème que nous avons soulevé : le fait qu'il faille préciser que les noms canoniques ne s'appliquent qu'à un langage propre au locuteur montre bien *a contrario* que ce qu'ils désignent n'est pas une simple combinaison de signes, mais un énoncé interprété, c'est-à-dire accompagné de sa signification. Autrement dit, la référence des noms canoniques n'est pas simplement une forme syntaxique mais pointe en fait vers le contenu sémantique de l'énoncé nominalisé. Et c'est là le point crucial qui jette un doute sur la nature purement syntaxique de l'opération de nominalisation canonique.

Alors, qu'est-ce qui est présent derrière la référence d'un nom canonique ? S'agit-il de l'énoncé considéré uniquement comme une suite de symboles indépendamment de sa signification, ou bien est-ce le contenu sémantique propre à cet énoncé. Nous serions tentés de répondre : les deux ! Considérez les assertions suivante :

1. L'énoncé «$_c$ la neige est blanche $_c$» est vrai
2. L'énoncé «$_c$ la neige est blanche $_c$» est composé de quatre mots (un article défini, un nom commun, un verbe et un adjectif)
3. L'énoncé «$_c$ la neige est blanche $_c$» nous dit quelque chose concernant la couleur d'une certaine forme de précipitation constituée de glace cristallisée et agglomérée en flocons pouvant être ramifiés d'une infinité de façons.

111. Plus précisément, la notation : $a \equiv \text{«}_c\ e\ _c\text{»} \pmod{trad.}$, signifie que l'énoncé d'un langage étranger dénoté par « a » se traduit dans notre idiolecte, modulo une traduction appropriée, par l'énoncé désigné par le nom canonique « «$_c$ e $_c$» ».

4. L'énoncé «$_c$ la neige est blanche $_c$» est le plus fameux exemple de Tarski.

Un locuteur francophone compétent saisira sans trop de difficulté le sens de ces assertions.[112] Et c'est au fond assez remarquable. Ici, l'énoncé dénoté par le nom canonique est à la fois identifié comme construction syntaxique (2), mais également comme porteur d'un certain contenu sémantique (3), tout en se voyant attribué la propriété d'être vrai (1) et en étant relié à une description définie qui en précise la provenance historique (4). Sans doute une théorie de la citation satisfaisante doit pouvoir rendre compte de *tous* ces usages. Mais cela nous mène un peu loin des règles minimales. Disons simplement que l'analyse de Rivenc conduit à penser que la nominalisation,[113] à l'oeuvre dans les T-équivalence, ou dans les règles minimales, s'appuie sur une compréhension proche de celle exhibée dans (3).[114] Sans chercher à déterminer plus avant ce que sera une « bonne » théorie des guillemets,[115] nous voudrions simplement tirer les leçons de tout ceci pour la question de la logicité de la vérité.

Malgré l'apparente trivialité avec laquelle nous maîtrisons l'usage des expressions guillemetées, tout porte à croire qu'il s'agit de phénomènes linguistiques très complexes et que les mécanismes qui les régissent dépassent de loin les ressources de la logique, telle qu'elle est caractérisée par l'approche inférentielle. François Rivenc achève son raisonnement en concluant que le déflationnisme, du moins dans sa version décitationnelle, réfute l'hypothèse qui le fonde : la théorie de la décitation repose sur une mauvaise analyse de la citation.[116] Dans le cadre de notre discussion, étant donné les ressources qui semblent

112. Brabanter, 2005 contient de nombreux et amusants exemples de ce type.
113. ou la citation, l'autonymie, ... quelle que soit l'appellation que l'on choisira au final.
114. Peut-être aussi pourrions-nous voir dans les expressions :

«$_c$ A $_c$» est vrai

des cas un peu bizarres de citations mixtes. Les citations mixtes, ou hybrides, sont des cas particuliers de citations où une expression *semble* à la fois mentionnée et utilisée. L'intérêt et l'importance des citations mixtes pour une théorie des guillemets ont été pour la première fois mis en avant dans la littérature philosophique par (Davidson1979) —du moins dans la sphère anglo-saxonne : (Brabanter2005) nous rappelle que Davidson a eu de nombreux précurseurs que la postérité n'a pas retenus—, avec son célèbre et malicieux exemple :

> Quine déclare que la citation « ... présente un certain caractère de bizarrerie ». (Davidson, 1979, p. 28)

L'idée que les noms canoniques seraient des sortes de citations mixtes nous semble naturellement en accord avec les remarques de Rivenc. Il serait intéressant de poursuivre cette vue. Mais nous en remettons toutefois l'analyse plus poussée à un travail ultérieur.

115. Nous laissons ce soin aux linguistes et philosophes du langage spécialistes de cette question et leur souhaitons, au passage, bien du courage.
116. *Cf.* Rivenc, 2004, p 521.

nécessaires pour rendre compte de la nominalisation canonique, nous interprétons nous aussi les remarques précédentes comme apportant, non pas une preuve définitive,[117] mais du moins un indice très probant du fait que la nominalisation à l'oeuvre dans les règles minimales n'est probablement pas, ou pas seulement, une opération syntaxique. *A fortiori*, elle semble encore moins être une opération purement logique. Si cette analyse est correcte, alors malgré les apparences, sans doute dues au fait que, graphiquement, l'énoncé est contenu dans la notation de son nom canonique, les règles minimales sont (vraiment très) loin d'être purement structurelles. Et, s'il y a plus dans la construction et la maîtrise des noms canoniques qu'une manipulation de symboles, il est même douteux de qualifier la vérité d'outil logico-syntaxique : même dans le cas des règles minimales, non seulement chacune de leurs instances s'articule sur une connaissance de la référence du nom canonique qui y figure, mais la nominalisation qui permet de construire ces noms canoniques est telle, et *doit* être telle,[118] que chaque nom « «$_c$ A $_c$» » porte avec lui le contenu de « A ».

2.5 Conclusion du chapitre

Il est temps de faire le point. Pour essayer de donner un sens précis à la thèse déflationniste selon laquelle la vérité serait une « sorte de notion logique », nous avons proposé un critère de démarcation emprunté aux approches inférentielles ou preuve-théoriques de la logicité. Pour pouvoir appliquer ce critère, nous avons reformulé les axiomatisations déflationnistes de la vérité sous la forme de règles d'inférence. Nous suivions en cela une méthode d'analyse de la vérité déflationniste déjà développée par Harold Hodes puis plus récemment par Henri Galinon. Le critère inférentiel de logicité se compose de trois conditions : la structuralité (pure) des règles à laquelle vient s'ajouter une double exigence d'harmonie globale et locale. Comme l'a souligné Hodes, les règles pour « vrai » nécessitent d'introduire des noms ou des termes auxquels s'applique le prédicat de vérité. Ceci a pour conséquence que les règles pour la vérité ne remplissent ni

117. L'absence actuelle, à notre connaissance, de théorie purement syntaxique des citations satisfaisante ne peut suffire à établir l'impossibilité de l'existence d'une telle théorie. Peut-être que reste à découvrir une théorie précise et purement syntaxique de la nominalisation canonique qui permettrait de répondre aux doutes que nous avons soulevés quant à la logicité des règles minimales. S'il faut en croire les diverses théories de la citation actuellement disponibles, cette perspective semble néanmoins extrêmement réduite. Un simple coup d'oeil à Cappelen et Lepore, 2012 ou à Brabanter, 2005 devrait suffire à s'en convaincre.

118. pour que les règles minimales soient valides.

la condition de structuralité, ni celle d'harmonie globale : elles ne sont pas conservatives sur la logique pure et elles s'appuient sur une connaissance sémantique préalable. Ceci semble disqualifier la vérité en tant que notion logique.

Face à cette difficulté, Henri Galinon a proposé une défense intéressante de la position déflationniste. En introduisant des noms canoniques primitifs, notés au moyen de guillemets, et en réduisant la signification de « vrai » à des règles minimales formulées à partir de ces noms, il a pu argumenter en faveur de la nature quasi logique de la notion vérité telle qu'elle est comprise par les déflationnistes : les règles minimales ne satisfont pas à strictement parler le critère de logicité, mais il ne s'en faut que très peu et la vérité déflationniste s'appuie sur des ressources particulièrement faibles.

Cependant, cette démonstration ne nous semble pas véritablement probante. Pour que logicité inférentielle et vérité déflationniste se rencontrent, l'argument nécessite à la fois d'aménager le critère de logicité (*exit* l'harmonie globale) et d'accepter des hypothèses supplémentaires (l'existence de noms canoniques primitifs). Mais ce fragile équilibre ne tient pas. Tout d'abord, l'existence de noms canoniques remplissant bien la mission qui leur est ici attribuée nous semble loin d'être aussi innocente et recevable qu'il peut paraître à première vue. En outre, les traiter comme des primitifs s'accorde mal avec l'approche inférentielle de la signification selon laquelle les règles doivent mettre au jour tous les mécanismes qui président à l'usage d'un concept. Lorsque l'on s'interroge un peu plus sur la nature de ces noms canoniques, ce « quelque chose comme ce que l'on obtient par une mise entre guillemets », il apparaît que leur construction et la maîtrise de leurs références mettent en jeu des processus complexes, au point que non seulement ces processus dépassent les seules ressources de la logique mais qu'il même douteux qu'on puisse les réduire à une simple opération syntaxique. À cette principale difficulté, s'ajoutent, d'une part, le fait que les règles minimales ne sont conservatives que sur une théorie de la syntaxe —il nous faut donc oublier ou modifier substantiellement le critère d'harmonie globale— et, d'autre part, le fait que le rôle expressif du prédicat de vérité semble remettre en cause la réduction de la signification de « vrai » aux seules règles minimales.

Bien sûr, ces doutes et ces remarques ne constituent pas une réfutation sans appel de la quasi logicité des règles minimales. Il faut bien admettre que rien n'interdit de *postuler* que des noms canoniques transparents, immédiatement donnés et primitifs existent dans le cerveau ou l'esprit des locuteurs humains qui emploient le prédicat de vérité. Rien

n'empêche non plus de partir de cette hypothèse « auxiliaire » et de l'associer à une lecture libérale et affaiblie du critère de logicité proposé par Dummett et Prawitz afin d'argumenter en faveur de la « quasi » logicité de la notion déflationniste de vérité. Après tout, il n'y a pas de lois contre ça. Une telle analyse montre-t-elle la cohérence des positions déflationnistes, renforçant par là leur plausibilité, ou n'obtient-on en conclusion que ce que l'on a bien voulu mettre dans nos hypothèses ? L'argument est-il vraiment convaincant ou ne peut-il servir qu'à prêcher les convertis ? Sur ce point, chacun pourra sans doute se faire son opinion personnelle. Pour les raisons que nous avons exposées, nous penchons pour la seconde alternative : même en acceptant que la signification de « vrai » est entièrement déterminée par des règles d'introduction et d'élimination, on ne peut pas sérieusement qualifier la vérité de notion logique à la lumière du critère inférentiel de logicité. Lorsque les déflationnistes nous disent que la vérité est « une sorte de notion logique », nous ne pouvons donc y voir qu'une manière métaphorique et rélâchée de parler peut-être plus trompeuse qu'éclairante.

2.6 Appendice technique

Pour ne pas alourdir la lecture, nous avons regroupés dans cette section, les résultats techniques auxquels il est fait référence dans la discussion qui précède.

2.6.1 Systèmes en Déduction Naturelle

Nous présentons tout d'abord un système de déduction naturelle à la Prawitz.

Logique Positive (LP) :

Un premier groupe de règles donne la Logique Positive.

- Règles pour la conjonction :

$$\wedge\text{-Intro}\ \frac{A \qquad B}{A \ \wedge B} \qquad \frac{A \wedge B}{A}\ \wedge\text{-Elim (G)} \qquad \frac{A \wedge B}{B}\ \wedge\text{-Elim (D)}$$

- Règles pour la disjonction :

$$\vee\text{-Intro (G)}\ \frac{A}{B \ \vee A} \qquad \frac{A}{A \ \vee B}\ \vee\text{-Intro (D)} \qquad \frac{A \vee B \qquad \begin{matrix}[A]\\ \vdots \\ C\end{matrix} \qquad \begin{matrix}[B]\\ \vdots \\ C\end{matrix}}{C}\ \vee\text{-Elim}$$

- Règles pour l'implicaton :

$$\rightarrow\text{-Intro}\ \frac{\begin{matrix}[A]\\ \vdots \\ B\end{matrix}}{A \rightarrow B} \qquad \frac{A \qquad A \rightarrow B}{B}\ \rightarrow\text{-Elim}$$

- Règles pour le quantificateur existentiel :

$$\exists\text{-Intro}\ \frac{A(t)}{\exists x A(x)} \qquad \frac{\exists x A(x) \qquad \begin{matrix}A[a]\\ B\end{matrix}}{B}\ \exists\text{-Elim}$$

- Règles pour le quantificateur universel :

$$\forall\text{-Intro}\ \frac{A(t)}{\forall x A(x)} \qquad\qquad \frac{\dfrac{\forall x A(x)}{B}}{A(t)}\ \forall\text{-Elim}$$

La Logique Positive ne contient pas de négation.

Logique Minimale (LM) :

En ajoutant les règles suivantes pour la négation, on obtient la Logique Minimale :

- Règles pour la négation :

$$\neg\text{-Intro}\ \frac{\begin{array}{c}[A]\\ \vdots\\ \bot\end{array}}{\neg A} \qquad\qquad \frac{A \qquad \neg A}{\bot}\ \neg\text{-Elim}$$

Logique Intuitionniste (LI) :

À l'ensemble des règles ci-dessus, qui donnent la logique minimale, on ajoute la règle suivante pour obtenir la logique intuitionniste :

- Ex falso Quod Libet :

$$\bot\text{-Elim}\ \frac{\bot}{A}$$

Logique Classique (LC) :

Pour obtenir la logique classique, on ajoute aux règles pour la logique intuitionniste l'une des deux règles ci-dessous :

- Loi du Tiers-exclu :

$$\text{LTE}\ \frac{}{A \vee \neg A}$$

• Elimination des Doubles Négations :

$$\frac{\neg\neg A}{A}\ \neg\neg\text{-Elim}$$

2.6.2 Les règles pour « Vr » ne sont pas conservatives sur LM $\cup \{=\}$

Nous reprenons ici une démonstration de HALBACH, 2001b que nous adaptons à notre cadre de discussion.

Soit **S** notre système de départ : un langage $\mathcal{L}$ pour la logique du premier ordre auquel s'ajoute un symbole pour l'égalité « = ». Comme système de preuve, nous prenons la logique minimale **LM**, dont nous avons rappelé ci-desssus les règles habituelles en déduction naturelle. A ce système de règles, on ajoute des règles pour l'égalité . En particulier, **S** contient un principe de substitution des identiques (PSI) : pour tous termes a, b du langage, pour tout symbole de prédicat P, la règle

$$\text{PSI}\ \frac{a = b}{P(a) \leftrightarrow P(b)}$$

est acceptable. On étend à présent, notre système logique au moyen des règles pour Vr (et de noms pour les énoncés de $\mathcal{L}$). On obtient **S'** avec pour langage est $\mathcal{L}'$, muni des règles

$$\textbf{LM} \cup \{\text{Règles pour l'égalité}\} \cup \{Vr\text{-Intro}, Vr\text{-Elim}\}.$$

Dans **S'**, on peut dériver l'énoncé $\exists x \exists y(x \neq y)$, qui est un énoncé de notre langage d'origine $\mathcal{L}$ non dérivable dans **S**.

Démonstration. Considérons l'énoncé « $A \to A$ », qui se prouve inconditionnellement dans **LM** par modus ponens. Soit a un nom de cet énoncé et soit b un nom de la négation de cet énoncé : « $\neg(A \to A)$ ».
Dans **S'** on peut dériver $Vr(a)$:

$$\frac{A \to A}{Vr(a)}\ \text{Vr-Intro}$$

On peut également réfuter $Vr(b)$:

$$\dfrac{\dfrac{\dfrac{[Vr(b)]^1}{\neg(A \to A)}\ \text{Vr-Elim} \qquad A \to A}{\bot}\ \neg\text{-Elim}}{\neg Vr(b)}\ \neg\text{-Intro, 1}$$

Dès lors,

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[a=b]^1}{Vr(a)\leftrightarrow Vr(b)}\text{PSI} \quad Vr(a)}{Vr(b)} \quad \neg Vr(b)}{\dfrac{\bot}{a\neq b}\,\neg\text{-Intro, 1}}\,\neg\text{-Elim}}{\exists y(a\neq y)}\,\exists\text{-Intro}}{\exists x\exists y(x\neq y)}\,\exists\text{-Intro}$$

Ainsi, **S'** n'est pas une extension conservative de **S** [119]. □

2.6.3 Les règles pour « Vr » sont conservatives sur la syntaxe

Le preuve ci-dessous est adaptée de Halbach (2014, p. 55-56).

Soit $\mathcal{L}$ un langage du premier ordre et soit T une théorie quelconque [120] exprimée dans $\mathcal{L}$. Supposons qu'on étende T par une théorie $T \subseteq T'$ de la syntaxe de $\mathcal{L}$. Plus précisément, pour chaque énoncé ϕ de $\mathcal{L}$, T' contient un nom de cet énoncé, noté $\overline{\phi}$ [121].

Supposons à présent qu'on étende T' par une théorie déflationniste de la vérité. Pour ce faire, on étend $\mathcal{L}$ par un nouveau symbole de prédicat « Vr » et on ajoute à T' la collection des T-équivalences : $T'_{Vr} := T' \cup \{Vr(\overline{\phi}) \leftrightarrow \phi \mid \phi \in \mathcal{L}\}$. Montrons que T'_{Vr} est conservative sur T' :

Démonstration. Soit π une preuve dans T'_{Vr} d'un énoncé ψ de $\mathcal{L}$: $T'_{Vr} \vdash_\pi \psi$.

Dans cette preuve apparaissent au plus un nombre fini d'axiomes de la forme

$$Vr(\overline{\phi_i}) \leftrightarrow \phi_i \quad 1 \leq i \leq n.$$

On définit dans T', $\Phi(x) := (x = \overline{\phi_1} \wedge \phi_1) \vee (x = \overline{\phi_2} \wedge \phi_2) \vee \ldots \vee (x = \overline{\phi_n} \wedge \phi_n)$

Pour chaque $1 \leq i \leq n$, T' prouve

119. On notera que la preuve fonctionne pour la logique minimale avec égalité et donc, *a fortiori*, pour les systèmes correspondants en logique intuitionniste ou classique.

120. T peut être une théorie ne contenant que la logique pure, c'est-à-dire l'ensemble des validités de la logique pure (l'ensemble des théorèmes de **LM**, **LI** ou encore **LC** selon le système considéré).

121. $\overline{\phi}$ peut être $\ulcorner\phi\urcorner$ si T' contient une théorie arithmétique telle que $\mathcal{Q}$ ou $\mathcal{PA}$. Mais nous ne voulons pas nous limiter aux cas où la syntaxe de $\mathcal{L}$ est donnée par un codage dans l'arithmétique. T' pourrait aussi bien contenir une théorie de la concaténation et des noms pour chaque symboles primitif de $\mathcal{L}$, T' pourrait contenir les ressources nécessaires pour donner des descriptions structurelles à la Tarski de chaque énoncé de $\mathcal{L}$, ou bien encore T' pourrait nous donner les ressources nécessaires pour mettre entre guillemets chaque énoncé de $\mathcal{L}$.

$$\Phi(\overline{\phi_i}) \leftrightarrow \phi_i.$$

Dans π, on remplace toutes les occurrences de « Vr » par « Φ ». Chaque axiome $Vr(\overline{\phi_i}) \leftrightarrow \phi_i$ qui apparaissait dans π est donc transformé en $\Phi(\overline{\phi_i}) \leftrightarrow \phi_i$. Devant chacune des formules $\Phi(\overline{\phi_i}) \leftrightarrow \phi_i$ apparaissant dans notre transformation de π, on place une preuve de cette formule dans T'. On obtient ainsi une preuve de ψ dans T'. (*NB :* ψ est un énoncé de $\mathcal{L}$ ne contenant pas d'occurrence de « Vr », et par conséquent il est laissé intact par notre transformation de π).

□

Deuxième partie

Vérité et conservativité

Chapitre 3

Les termes du débat

Dans ce chapitre nous présentons et discutons un argument contre les théories déflationnistes avancé, de manière indépendante, par Stewart Shapiro (Shapiro, 1998b) et Jeffrey Ketland (Ketland, 1999). L'argument consiste à placer le déflationniste face à deux contraintes d'adéquation, qui semblent s'imposer à une théorie de la vérité déflationniste, mais qui s'avèrent incompatibles. Le raisonnement de Shapiro et Ketland s'appuie sur certains éléments techniques de logique, au premier rend desquels la notion de conservativité. Ces deux auteurs affirment en effet qu'une contrainte de conservativité pèse de manière essentielle sur un prédicat de vérité déflationniste. C'est la première contrainte. Par ailleurs, Shapiro et Ketland considèrent qu'une théorie de la vérité satisfaisante doit être réflexive. C'est-à-dire qu'une théorie de la vérité adéquate doit pouvoir rendre compte de l'emploi qui est fait du prédicat « vrai » dans certains arguments sémantiques. Plus précisément, lorsqu'on ajoute à une théorie de base $\mathcal{S}$ des axiomes pour la vérité, on doit pouvoir être en mesure de formaliser un argument inductif permettant d'établir l'énoncé : « tous les théorèmes de $\mathcal{S}$ sont vrais ». C'est la seconde contrainte. Or, on sait depuis les résultats d'incomplétude de Gödel que dès lors que $\mathcal{S}$ contient un minimum d'arithmétique, toute théorie étendant $\mathcal{S}$ et permettant de prouver que tous les théorèmes de $\mathcal{S}$ sont vrais sera non conservative. Ketland et Shapiro en concluent que cette double exigence de conservativité et de réflexivité place le déflationniste devant un dilemme apparemment insoluble.

Le squelette de l'argument peut donc s'énoncer de manière simple et compacte comme suit :

(1) Une théorie déflationniste de la vérité doit être conservative.

(2) Une théorie adéquate de la vérité doit être réflexive.

(3) Toute théorie de la vérité réflexive sera non conservative.

Donc,

(C) Les théories déflationnistes de la vérité sont inadéquates. [1]

Depuis sa parution, cet « argument de la conservativité [2] » ou « argument de la réflexion [3] », ainsi qu'il est parfois intitulé, a fait couler beaucoup d'encre. Pour en contester la force, le déflationniste peut choisir de rejeter l'une ou l'autre des prémisses. Ainsi, par exemple, Halbach (Halbach, 2001b,c) ou Horwich (communication personnelle, citée dans (Tennant, 2010, p. 439)) rejettent (1), Tennant (Tennant, 2002, 2010, 2005) rejette (2), Field (Field, 1999) accepte (1) et (2) mais fournit une autre explication de (3) ...

Dans ce qui suit nous reprenons les principaux points de cette querelle toujours en cours. Tout d'abord nous introduisons les deux contraintes proposées par Shapiro et Ketland (section 1 & 2). Puis nous exposons les principaux résultats techniques sur lesquels s'appuient la discussion de leur argument (section 3). Une fois rappelé ces prérequis, nous examinons dans le chapitre suivant les diverses voies qui ont été explorées autour de l'argument de la conservativité par les auteurs déflationnistes ou leurs détracteurs.

3.1 La conservativité : critère de l'absence de « substantialité »

Au delà de la diversité des formulations qu'ils adoptent, l'une des thèses centrales communes aux auteurs déflationnistes semble être que la vérité ne saurait jouer de rôle explicatif au sein de nos discours sur le monde. La vérité ne serait pas une propriété « substantielle », elle serait métaphysiquement « sans poids » [4], le prédicat de vérité ne serait qu'un outil linguistique permettant d'augmenter le pouvoir expressif de notre langage ; on ne devrait donc pas s'attendre à ce que l'attribution de vérité à tel ou tel énoncé ou à telle ou telle proposition permette d'expliquer tel ou tel phénomène. Ketland et Shapiro considèrent que de cet assemblage d'idées découle une contrainte de conservativité,

1. Nous empruntons cette formulation à Armour-Garb, 2012, p. 261-262
2. Selon la terminologie de Field, 1999.
3. Selon la terminologie de Ketland, 2005.
4. Shapiro parle de vérité « métaphysiquement maigre » (*metaphysically thin*, Shapiro, 1998a) à propos de la conception que s'en font les déflationnistes.

qui s'impose aux théories déflationnistes. À bien des égards, une contrainte de conservativité peut effectivement apparaître comme une traduction naturelle de l'absence de rôle explicatif de la vérité et comme une manière de donner un sens technique précis à l'idée que la vérité ne peut tenir de tel rôle. Historiquement, la recherche de résultats de conservativité a en effet très souvent été associée à des programmes réductionnistes visant à « dégonfler » certains concepts et à justifier un emploi purement instrumental de certaines méthodes jugées métaphysiquement « douteuses ». Son emploi en philosophie de la logique et des mathématiques a été introduit par Hilbert comme l'un des éléments centraux de son fameux Programme. Avant d'expliquer et d'illustrer quelque peu comment la conservativité a pu être employée en ce sens, commençons par définir précisément ce dont il s'agit.

3.1.1 Définitions

La notion de conservativité est une propriété technique de logique mathématique qui s'applique à des théories formalisées. Sous une première forme, cette propriété s'énonce comme suit :

Définition 5. CONSERVATIVITÉ DÉDUCTIVE :
Soient $\mathcal{L} \subseteq \mathcal{L}'$ deux langages formels et soient $T \subseteq T'$ deux théories exprimées respectivement dans $\mathcal{L}$ et dans $\mathcal{L}'$. T' est une extension (déductivement) conservative de T si et seulement si

$$\forall \varphi \in \mathcal{L}, T' \vdash \varphi \Rightarrow T \vdash \varphi.$$

En d'autres termes, T' est une extension conservative de T si toute assertion formulée dans $\mathcal{L}$, le langage de T, démontrable à partir de T' est déjà démontrable (quoiqu'éventuellement de manière beaucoup moins rapide et pratique) dans T. Le passage de T à T' laisse notre stock de $\mathcal{L}$-théorèmes inchangé, et n'améliore pas nos capacités explicatives vis-à-vis des faits énoncés dans $\mathcal{L}$.

C'est sous cette forme, c'est-à-dire dans une version déductive (on dit aussi syntaxique) que la conservativité a été introduite par Hilbert. Il existe cependant une seconde version de la propriété de conservativité qu'on peut appeler version *sémantique* de la conservativité. Elle s'appuie non plus sur la notion de démonstration (comme enchaînement fini d'énoncés, obtenu suivant certaines règles effectives) mais sur celle de conséquence sémantique (φ est conséquence logique, au sens sémantique, de T si tout

modèle de T satisfait φ).[5] La définition de ce second type de conservativité est la suivante :

Définition 6. CONSERVATIVITÉ SÉMANTIQUE :
Soient $\mathcal{L} \subseteq \mathcal{L}'$ deux langages formels et soient $T \subseteq T'$ deux théories exprimeée respectivement dans $\mathcal{L}$ et dans $\mathcal{L}'$. T' est une extension (sémantiquement) conservative de T si et seulement si

$$\forall \varphi \in \mathcal{L}, T' \models \varphi \Rightarrow T \models \varphi.$$

On a bien là une version modèle-théorique de la conservativité : T' est sémantiquement conservative sur T si tout énoncé de $\mathcal{L}$ qui est conséquence logique de T' est déjà conséquence logique de T.

La distinction entre ces deux types de conservativité n'est pas toujours importante ou pertinente. D'après le théorème de complétude de Gödel, les deux notions de conservativité coïncident dans le cadre de la logique du premier ordre : toute extension syntaxiquement conservative sera sémantiquement conservative et vice versa. En revanche, il faut bien noter que ce ne sera plus nécessairement le cas si l'on quitte la logique du premier ordre. Au second ordre classique[6], il existe des théories T telles qu'on peut trouver un énoncé φ qui est conséquence logique (au sens sémantique du terme) de T mais n'est pas démontrable à partir de T par recours aux méthodes de dérivation standard[7]. C'est une conséquence directe de l'incomplétude syntaxique de la logique du second ordre et du caractère ineffectif de sa relation de conséquence sémantique. Dans un tel cadre, la contrainte de conservativité sémantique est moins forte que la contrainte de conservativité syntaxique : il existe des extensions sémantiquement conservatives qui ne sont pas syntaxiquement conservatives.[8] Lorsqu'on parle de conservativité dans la littérature

5. La notion de conséquence sémantique à laquelle il est fait ici référence est évidemment la notion classique issue de la théorie des modèles standard, où les modèles ont pour structure sous-jacente des ensembles, qui eux-mêmes peuvent être définis dans une métathéorie telle que ZFC. Nous ne voulons pas entrer ici dans les débats concernant la « bonne » notion de conséquence logique tels qu'ils ont pu être ouverts, à la suite de Tarski, par Etchemendy, Kreisel ou Boolos.

6. c'est-à-dire munie de la sémantique standard, où les variables du second ordre portent sur toutes les sous-parties du domaine et ne sont pas limitées à une sous collection propre de ces sous-parties, comme c'est le cas dans les modèles « généraux » de Henkin.

7. c'est-à-dire un système de preuve récursif.

8. Un exemple trivial pour fixer les idées : soient T une théorie du second ordre exprimée dans $\mathcal{L}$, et soit φ un énoncé de $\mathcal{L}$ tel que $T \models \varphi$ mais $T \nvdash \varphi$ (de tels T et φ existent dès lors que $\vdash$ désigne un système de preuve effectif). On pose $T' = T \cup \{\varphi\}$. Alors T' est sémantiquement conservative sur T : en effet comme $T \models \varphi$, pour tout $\psi \in \mathcal{L}$, $T' = T \cup \{\varphi\} \models \psi \Rightarrow T \models \psi$. Pour autant, T' *n'est pas* syntaxiquement conservative sur T : $T' = T \cup \{\varphi\} \vdash \varphi$, alors que, par hypothèse, $T \nvdash \varphi$.

logique ou philosophique, c'est généralement à l'une ou l'autre de ces formes ci-dessus qu'on veut faire référence. Dans ce qui suit, sauf précision contraire, nous parlerons de conservativité dans les deux sens ci-dessus, qui redisons-le, se confondent lorsqu'on se situe dans un cadre logique du premier ordre. Il existe cependant une troisième sorte de conservativité pouvant relier deux théories formelles. Ce troisième type de conservativité aura son importance lorsque nous discuterons, dans le chapitre suivant, l'un des axes de réponse que l'argument de Ketland (1999) et Shapiro (1998b) a suscités. Nous en donnons dès à présent la définition, bien que nous repoussions sa discussion plus détaillée au chapitre suivant :

Définition 7. Conservativité modèle-théorique

Soient T une théorie exprimée dans un langage $\mathcal{L}_T$ et T' une théorie étendant T exprimée dans un langage $\mathcal{L}_{T'} \supseteq \mathcal{L}_T$. On dit que T' est modèles-théoriquement conservative sur T ssi toute $\mathcal{L}_T$-structure qui est modèle de T peut être enrichie en une $\mathcal{L}_{T'}$-structure qui est modèle de T'.

Avant d'en venir à l'emploi de la conservativité au sein des débats touchant le déflationnisme en matière de vérité, nous donnons deux exemples historiques célèbres d'utilisation de cette notion (ou plutôt de *ces* notion*s*) en lien avec des entreprises de « déflation » ou de « dégonflement » de certains concepts. Ces exemples permettent non seulement d'illustrer des emplois typiques de la conservativité mais fournissent en outre des points de comparaison utiles pour évaluer la portée de la contrainte de conservativité censée s'imposer aux déflationnistes aléthiques d'après Ketland (1999) et Shapiro (1998b)

3.1.2 Conservativité, cohérence et instrumentalisme : le programe de Hilbert

La première utilisation de la conservativité dans le cadre d'un argument philosophique visant à « dégonfler » certaines notions remonte sans doute à Hilbert. Au tournant du vingtième siècle, la découverte des antinomies avait en effet plongé les mathématiques dans un état de crise : des principes de raisonnements jusque là considérés comme sûrs se révélaient susceptibles de mener à des contradictions. La mise au point de son fameux programme a constitué la réponse la plus élaborée de Hilbert à ce problème du

fondement des mathématiques. [9] Alors que certains mathématiciens et logiciens recommandaient de renoncer aux méthodes de raisonnement des mathématiques classiques, [10] Hilbert entendait asseoir l'ensemble des mathématiques sur une base solide au moyen d'une preuve de cohérence. Il s'agissait de montrer que les systèmes de preuves en usage au sein des mathématiques ne pouvaient aboutir à démontrer à la fois une proposition et sa négation. Par conséquent, il apparaissait nécessaire d'explorer non plus seulement les entités traditionnellement examinées par les mathématiciens mais de se donner les preuves elles-mêmes pour objet d'étude.

La première étape du Programme de Hilbert consiste donc à formaliser l'ensemble des raisonnements mathématiques. [11] Chaque proposition mathématique se verra traduite par une formule couchée dans un langage formel, et à chaque preuve correspondra une dérivation à l'intérieur d'un système axiomatisé. Une fois cette opération réalisée, il devient possible d'étudier mathématiquement les preuves. Les dérivations au sein d'un système axiomatisé peuvent en effet être vue comme de simples manipulations de suites de symboles au moyen de règles effectives de dérivation, abstraction faite du sens intuitif qui était originellement attribué aux énoncés. Les preuves sont alors considérées comme des opérations purement syntaxiques.

Dans un second temps, Hilbert distingue au sein des mathématiques une part élémentaire, qu'il appelle mathématiques « finitistes » [12], dont la cohérence et la correction sont considérées comme allant de soi. Cette part des mathématiques porte sur des objets quasi-concrets : les signes. Par exemple, les nombres considérés comme des suites (finies) de barres tracées sur le tableau, ou bien encore les combinaisons (finies) de symboles qui

9. Pour un exposé synthétique en français des principaux points du Programme de Hilbert, voir l'ultime chapitre de (Blanché et Dubucs, 1996, en particulier p. 364-370). En anglais, on pourra consulter également (Zach, 2009) ainsi que (Raatikainen, 2003) et (Mancosu, 1998, p 149-177).

10. Le finitisme de Kronecker ou l'école intuitionniste qui commençait à se développer autour de Brouwer sont de bons exemples d'un tel révisionnisme. C'est d'ailleurs en partie pour répondre à ces auteurs et à l'influence grandissante qu'ils exerçaient au sein de la communauté mathématique qu'Hilbert a développé son Programme. Et, lorsqu'il distingue les mathématiques « finitistes » (*cf.* ci-dessous), ou ayant contenu concret, réel, Hilbert espère aussi pouvoir répondre aux sceptiques en n'utilisant que leurs propres armes.

11. Sur ce point, Hilbert pouvait s'appuyer sur les travaux de Frege et de Russell. Les systèmes formels conçus par ces derniers arrivaient donc à point nommé.

12. Pour qualifier ces mathématiques « finitistes », Hilbert emploie aussi parfois les expressions « mathématiques réelles » ou « mathématiques concrètes », ou bien encore « mathématiques porteuses de contenu (*inhaltlich*) ». Le point crucial est que les énoncés de ces mathématiques « concrètes » sont des énoncés interprétés censés référer à des objets finis, contrairement aux formules des mathématiques « idéales » (voir ci-dessous) qui, elles, sont considérées et traitées comme de simples combinaisons de symboles dénuées de toute signification.

composent les formules de tel ou tel langage formel. En outre, les propriétés de base de ces objets, comme la succession ou la concaténation, nous sont directement accessibles. Les énoncés des mathématiques finitistes ont donc un contenu « réel » , « concret », et sont dotés d'une signification bien définie.

Cette part des mathématiques est tellement fondamentale qu'elle ne saurait être fondée ou réduite à autre chose. Elle s'impose à tous, y compris aux plus sceptiques, comme préalable indispensable à toute pensée rationnelle et à toute entreprise scientifique.Le passage suivant tiré de Hilbert, 1926, p. 228 illustre bien la conception de Hilbert [13] :

> « La condition préalable de l'application des inférences logiques et de l'effectuation d'opérations logiques est l'existence d'un donné dans la perception : à savoir l'existence de certains objets concrets extra-logiques qui en tant que sensations immédiates précèdent toute pensée. Pour que le raisonnement logique soit sûr, il faut que ces objets soient perçus dans toutes leurs parties et que leur occurrence, leur caractère distinct, leur succession ou leur juxtaposition se présentent à l'intuition en même temps que ces objets, comme quelque chose d'immédiat et qui ne se réduit pas ou n'a pas besoin d'être réduit à quoi que ce soit d'autre. Telle est la conception philosophique fondamentale qu'à mon sens exigent les mathématiques et d'ailleurs toute pensée, toute compréhension et toute communication scientifiques. En ce qui concerne particulièrement les mathématiques, l'objet de notre étude sera donc les signes concrets eux-mêmes dont nous savons, du point de vue que nous avons adopté, distinguer et reconnaître la forme. »

Hilbert n'a jamais pris la peine de tracer précisément la frontière de ces mathématiques « concrètes ». Néanmoins, on peut identifier ce sous domaine des mathématiques classiques avec ce qu'on appelle aujourd'hui l'arithmétique primitive récursive. [14] En termes contemporains, les énoncés « à contenu » des mathématiques finitistes correspondent à ce qu'on appelle aujourd'hui l'ensemble des formules Π^0_1 dans la hiérarchie arithmétique, c'est-à-dire les formules de la forme $\forall x_1 \forall x_2 \ldots \forall x_n \varphi(x_1, \ldots x_n)$ où φ est une formule sans

13. On en retrouve d'ailleurs des variantes ou des reprises quasi *verbatim* dans de nombreux textes de l'auteur (*cf.* Hilbert, 1931, 1922).

14. Cette interprétation s'est en grande partie imposée depuis Tait, 1981. Pour une défense assez récente de cette identification du finitisme hilbertien avec $\mathcal{PRA}$, voir aussi : Raatikainen, 2003 ; Tait, 2002. Notons, néanmoins, que la question de l'extension précise du domaine des mathématiques finitistes dans la lignée de Hilbert est toujours l'objet de débats entre spécialistes : outre les références déjà citées dans cette note, *cf.* Parsons, 1998, Detlefsen, 1990, Kreisel, 1958, 1970.

quantificateur [15] qui désigne une propriété primitive récursive. Ainsi, Hilbert inclut au sein des énoncés finitistes certaines formules arithmétiques universellement quantifiées. Ce point est délicat et mystérieux. Spontanément, on pourrait penser que, même si les entiers sont identifiés à des suites de signes, un énoncé contenant une quantification portant sur (tous) les entiers, c'est-à-dire, semble-t-il, portant sur une collection infinie d'objets, n'est pas acceptable d'un point de vue finitiste. Hilbert est peu disert à ce sujet. Mais en fait, Hilbert s'appuie sur une interprétation « prototypique » de la quantification universelle. Un énoncé du type $\forall x \varphi(x)$ doit se comprendre comme énonçant qu'il existe une méthode générique (et primitive récursive) permettant, pour chaque entier n particulier, fixé, de vérifier $\varphi(n)$:

> « [...]Le cas du quantificateur universel est moins clair. Hilbert semble penser que les énoncés de la forme "pour tout numeral a, $a+1 = 1+a$" sont finitistes. Cependant, il ne lit pas ceci comme une conjonction infinie mais "comme un jugement hypothétique qui revient à asserter quelque chose quand un numéral est donné." » Mancosu, 1998, p. 162 [16]

Cette interprétation permet, ou est censée permettre, de donner un sens à l'énoncé universel sans pour autant s'engager quant à l'existence d'une totalité infinie de nombres entiers, c'est-à-dire sans postuler l'existence d'un infini en acte. [17] La quantification existentielle *non bornée* est en revanche bannie. Selon Hilbert, elle n'est pas acceptable d'un point de vue strictement finitiste : les assertions d'existence qui ne peuvent être réduites à des disjonctions finies nous font sortir du domaine finitiste. En effet, lorsque nous sommes face à un énoncé du type $\exists x \varphi(x)$, nous n'avons *a priori* aucun moyen fini de déterminer si cette formule est vraie. Dès lors, quel sens finitiste pourrait-on attribuer à un tel énoncé ? Aucun selon Hilbert. [18]

Pour être exhaustif dans cette rapide description, signalons que les méthodes de preuves admissibles « au départ », [19] d'un point de vue strictement finitiste correspondent peu ou prou à un système équivalent à l'arithmétique de Robinson assortie d'un schéma

15. φ peut éventuellement contenir des quantifications *bornées*. Ces dernières pouvant être réécrites sous la forme de conjonctions ou de disjonctions *finies*.

16. Nous traduisons.

17. Pour une analyse minutieuse de la façon dont on peut donner un sens finitiste aux « propositions générales » de la forme $\forall x_1 \ldots \forall x_n \varphi(x_1, \ldots, x_n)$ sans supposer l'existence d'une infinité de nombres, nous renvoyons de nouveau à Tait, 1981.

18. Par ailleurs, il est clair que l'interprétation « prototypique » de la quantification universelle non bornée ne peut être adaptée au cas du quantificateur existentiel.

19. c'est-à-dire avant l'introduction des mathématiques idéales (*cf.* ci-dessous).

d'induction restreint aux seules formules sans quantificateur. [20] Il faut néanmoins signaler que l'arithmétique primitive récursive, $\mathcal{PRA}$, peut aussi être axiomatisée au moyen d'un langage sans quantificateur. C'est d'ailleurs sous cette forme qu'elle fut initialement introduite par Skolem (Skolem, 1923). Les « propositions générales » prennent alors la forme d'équations à variables libres. Par exemple, l'énoncé Π^0_1, $\forall x \forall y (x + y = y + x)$ s'écrira $x + y = y + x$. Sous cette forme $\mathcal{PRA}$, contient un symbole pour chaque fonction récursive primitive, qui sont alors caractérisées par des équations définitionnelles contenant des variables libres mais sans quantificateur. Si les deux formalisations sont équivalentes, il est clair que la forme sans quantificateurs sied particulièrment bien à l'interprétation « prototypique » des énoncés universels qu'Hilbert semble avoir eu en tête. Comme l'ont montré plus tard Curry, 1941 puis Goodstein, 1954, on peut même faire mieux : il est possible de donner une axiomatisation de $\mathcal{PRA}$ ne contenant aucun connecteur logique. Dans ce système les termes sont identifiés à des fonctions primitives récursives à zéro ou plusieurs variables. L'existence de tels systèmes apparait comme un élégante justification *a posteriori* des intuitions hilbertiennes selon lesquelles les mathématiques finitistes et les objets sur lesquels elles portent, forment un préalable incoutournable à toute pensée rationnelle et précèdent même la logique. [21]

Les mathématiques finitistes de Hilbert recèlent, on le voit, une dissymétrie en ce qui concerne la quantification. Plus généralement, elles présentent un désagréable manque d'uniformité : la classe des énoncés finitistes n'est pas stable par application des lois de la logique classique. Ainsi, la négation d'un énoncé finitiste n'est pas, en règle générale, un énoncé finitiste. [22] De même, un énoncé acceptable d'un point de vue finitiste peut compter parmi ses conséquences des énoncés inacceptables. En voici une illustration, due à Hilbert lui-même [23] : si p est un nombre premier donné, le procédé d'Euclide nous permet de montrer qu'il existe un nouveau nombre premier compris entre $p+1$ et $p!+1$. Cette dernière proposition contient une quantification existentielle bornée (par $p! + 1$) portant une propriété primitive récursive (être premier), elle est donc acceptable d'un

20. Pour une description détaillée d'une axiomatisation de $\mathcal{PRA}$ dans un formalisme contemporain, voir par exemple Simpson, 2009, p. 373-383.

21. Zach, 2009 insisite sur le fait que, pour Hilbert, les objets des mathématiques finitistes sont censés être antérieurs à toute logique.

22. La négation s'applique sans problème aux énoncés sans quantificateurs. Si $\varphi(x)$ est Δ^0_0, sa négation $\neg\varphi$ le sera également. En revanche, la négation d'un énoncé Π^0_1, $\forall x \varphi(x)$ est (équivalente à) un énoncé Σ^0_1, $\exists x \neg\varphi(x)$, c'est-à-dire à un énoncé possédant une quantification existentielle *non* bornée. L'emploi de la négation (classique) nous fait donc alors sortir du domaine des mathématiques finitistes.

23. *Cf.* Hilbert, 1926, p. 230.

point du vue finitiste. [24] Malheureusement, elle a aussi pour conséquence l'assertion qu'il existe un nombre premier strictement supérieur à p. Cette dernière assertion, quoiqu'apparemment plus faible que la précédente, est une affirmation existentielle non bornée, et n'est donc plus acceptable du point de vue finitiste.

Malgré cela, Hilbert n'entend pas renoncer à la logique classique et aux méthodes de raisonnement habituellement acceptées en mathématiques. Il propose donc de « prolonger » ou de « compléter » les mathématiques finitistes en introduisant ce qu'il appelle les mathématiques « idéales ». Aux énoncés finitistes à contenu s'ajoutent des énoncés « idéaux » qui permettent de rétablir l'emploi les règles habituelles de la logique. Pour autant, ces énoncés « idéaux » n'ont pas le même statut que les énoncés des mathématiques finitistes. Si ces derniers ont bien un contenu, les énoncés « idéaux » ne sont que de simples « manières de parler », des suites de symboles sans signification qu'il est commode d'introduire pour pouvoir simplifier le formalisme. Hilbert compare ce prolongement par des énoncés « idéaux » à la démarche, courante en mathématiques, qui consiste à introduire des objets imaginaires pour faciliter les calculs ou unifier des lois générales. [25]

Mais, si l'introduction des mathématiques idéales semble souhaitable sur un plan pratique, encore faut-il qu'elles ne nous induisent pas en erreur. Il convient de s'assurer que le passage des mathématiques finitistes à un système étendu au moyen d'énoncés idéaux ne va pas nous conduire à des contradictions. À ce stade, le Programme de Hilbert se décline sous deux formats.

24. Comme le remarque Hilbert, on peut d'ailleurs l'interpréter comme l'abréviation d'une disjonction finie.

25. Sur le prolongement par des énoncés idéaux en vue de retrouver les lois de la logique classique et sa comparaison avec la méthode des objets idéaux en mathématique :

> « De même que $i = \sqrt{-1}$ a été introduit pour conserver aux lois de l'algèbre, par exemple à celles qui portent sur l'existence et le nombre des racines d'une équation, leur forme la plus simple ; de même que l'introduction des facteurs idéaux est intervenue afin de maintenir les lois de la divisibilité pour les nombres entiers algébriques (par exemple lorsque nous définissons pour les nombres 2 et $1+\sqrt{-5}$ qui n'ont pas de diviseur commun réel, un diviseur commun idéal), il nous faut ici *ajouter les propositions idéales aux propositions finitistes* afin de maintenir en vigueur les règles f ormellement simple de la logique d'Aristote. » Hilbert, 1926, p. 231

Sur la différence de statut entre les énoncés à contenu des mathématiques finitistes et les énoncés idéaux :

> « Si nous généralisons cette conception, les mathématiques deviennent un réservoir de formules qui contiendra en premier lieu celles auxquelles correspond la communication de propositions finitistes, et ensuite d'autres formules qui n'ont pas de sens et qui constituent les *objets idéaux de notre théorie* » Hilbert, 1926, p. 232

Sous une première forme, on obtient un *Programme de Consistance*. La sûreté de l'ensemble des mathématiques étendues sera garantie en en donnant une preuve de cohérence. Soit $\mathcal{F}$ une théorie axiomatisée formalisant le fragment finitiste des mathématiques et soit $\mathcal{F} \cup \mathcal{I}$ une axiomatisation de l'ensemble des mathématiques étendues par les énoncés idéaux. Il s'agit alors de *démontrer* que $\mathcal{F} \cup \mathcal{I}$ est cohérente. De plus, on exigera que cette preuve soit obtenue en ne recourant qu'à des concepts et des méthodes finitistes. Il est d'ailleurs remarquable que la cohérence de $\mathcal{F} \cup \mathcal{I}$ puisse s'exprimer au moyen d'un énoncé finititste. En effet, par définition, une théorie est cohérente si elle ne contient pas de preuve débouchant sur une contradiction. Mais, pour une théorie formalisée, une preuve n'est rien d'autre qu'une dérivation, c'est-à-dire une suite *finie* de symboles, où chaque ligne contient un axiome ou une formule obtenue au moyen de règles *calculables* à partir des lignes précédentes, la conclusion se trouvant en dernière position. Par conséquent les preuves formalisées sont des objets acceptables et manipulables à partir des seules ressources finitistes. Plus précisément, on peut définir dans $\mathcal{F}$ une propriété *primitive récursive* $Pr(x, y)$ exprimant que x est une preuve de y *dans* $\mathcal{F} \cup \mathcal{I}$.[26] La cohérence de $\mathcal{F} \cup \mathcal{I}$ s'énonce alors par un énoncé Π^0_1, à savoir, $Con(\mathcal{F} \cup \mathcal{I}) :=_{def} \forall x \neg Pr(x, \bot)$ où $\bot$ désigne une contradiction.[27] Le Programme de Consistance sera donc réalisé si on parvient à établir que

$$\mathcal{F} \vdash Con(\mathcal{F} \cup \mathcal{I}).^{28}$$

Parallèlement, il existe aussi une seconde version du Programme de Hilbert, parfois appelée *Programme de Conservatitivité*. L'inocuité du passage de $\mathcal{F}$ à $\mathcal{F} \cup \mathcal{I}$ est ici établie par un résultat de conservativité (déductive) :

$$\text{Pour tout énoncé } \varphi \text{ du langage de } \mathcal{F},\ \mathcal{F} \cup \mathcal{I} \vdash \varphi \Rightarrow \mathcal{F} \vdash \varphi^{29}$$

26. Sous réserve, bien sûr, que $\mathcal{F} \cup \mathcal{I}$ soit une théorie primitive récursive. Sous cette hypothèse, on peut définir $Pr(x, y)$ soit en considérant directement les objets finis et quasi-concrets de la syntaxe de $\mathcal{F} \cup \mathcal{I}$, *i.e.* les suites de symboles appropriées elles-mêmes ; soit en ayant recours à un codage dans une arithmétique élémentaire acceptable d'un point de vue finitiste. Auquel cas, $Pr(x, y)$ doit se lire : x est le code d'une dérivation dans $\mathcal{F} \cup \mathcal{I}$ de la formule codée par y. Il est en tout cas clair que les ressources des mathématiques finitistes (qui contiennent les propriétés primitives récursives) suffisent à définir $Pr(x, y)$.

27. Par exemple, pour fixer les idées, $\bot$ désigne l'énoncé « $0 \neq 1 \wedge 0 = 1$ ». L'énoncé « $\forall x \neg Pr(x, \bot)$ » est alors bien une « proposition générale » portant sur une propriété récursive. Cet énoncé est donc bien dôté d'un contenu « réel » et relevant de $\mathcal{F}$. L'importance cruciale pour Hilbert d'intégrer au sein des mathématiques finitistes les énoncés Π^0_1 apparaît ici en pleine lumière. Ce n'est qu'à cette condition que la cohérence des mathématiques étendues peut être exprimée et appréhendée à l'intérieur du seul fragment finitiste.

28. Ou peut-être, étant donné que la cohérence de $\mathcal{F}$ est considérée comme allant de soi et qu'elle s'impose à tous : $\mathcal{F} + Con(\mathcal{F}) \vdash Con(\mathcal{F} \cup \mathcal{I})$. *Cf.* (Raatikainen, 2003).

29. C'est bien sous cette forme *déductive* qu'Hilbert entendait la notion de conservativité. De toute

En montrant que $\mathcal{F} \cup \mathcal{I}$ est conservative sur $\mathcal{F}$, on obtient la garantie que toute preuve d'un énoncé ayant un contenu réel peut, du moins en théorie, être expurgée de tout élément idéal. Les éléments idéaux ne sont bien que des « manières de parler », de simples « détours » qui sont peut-être utiles en pratique, puisqu'ils simplifient nos méthodes de démonstration et facilitent les calculs, mais qui n'augmentent en rien nos capacités de preuves concernant les énoncés « réels ». [30] Là encore, il semble naturel d'exiger en outre que ce résultat de conservativité soit établi par des méthodes strictement finitistes. Le *Programme de Conservativité* permet d'envisager une justification purement heuristique de l'emploi de méthodes dépassant le cadre finitiste : leur inertie déductive vis-à-vis des mathématiques à contenu assure que l'on ne perd (ni ne gagne) rien à les adopter et autorise leur usage, tandis leur commodité nous commande d'y avoir recours. Selon la belle formule de Jean Largeault :

> « on élimine les concepts et les propositions transfinies *en théorie* pour avoir le droit de les conserver *en pratique*. » [31]

Ces deux versants du programme de Hilbert ne sont pas sans rapports l'un avec l'autre. Sous certaines hypothèses, ils sont même équivalents au sens où la réalisation de l'un entraine la réalisation de l'autre. Supposons par exemple que l'on dispose d'une preuve finitiste de la cohérence de $\mathcal{F} \cup \mathcal{I}$. Supposons en outre que $Pr(x, y)$, notre prédicat de prouvabilité dans $\mathcal{F} \cup \mathcal{I}$, satisfasse certaines conditions naturelles de dérivabilité, [32] on a alors le résultat suivant, dû à Kreisel :

Soit T une théorie idéale (*i.e.* telle que $\mathcal{F} \subseteq T$, l'inclusion pouvant être stricte), soit φ un énoncé Π^0_1, $T \vdash \varphi \Rightarrow \mathcal{F} + Con(T) \vdash \varphi$

Dès lors, si $\mathcal{F} \vdash Con(\mathcal{F} \cup \mathcal{I})$, le résultat ci-dessus se réduit à

$$\mathcal{F} \cup \mathcal{I} \vdash \varphi \Rightarrow \mathcal{F} \vdash \varphi, \text{ pour toute } \varphi \in \Pi^0_1,$$

ce qui est exactement la conservativité de $\mathcal{F} \cup \mathcal{I}$ pour les énoncés finitistes.

manière, à l'époque où il développait sont Programme, les outils de théorie des modèles nécessaires pour conceptualiser la notion sémantique de conservativité n'avaient pas encore été mis au point.

30. Cette seconde forme de Programme de Hilbert s'inscrit aussi dans l'idéal, cher au mathématicien, de pureté des méthodes selon lequel la preuve d'un énoncé relevant d'un certain domaine D des mathématiques ne doit faire intervenir que des éléments pris dans D.

31. (Largeault, 1972, p 215)

32. Il nous faut supposer que Pr satisfait les conditions de dérivabilité de Löb-Hilbert-Bernays. *Cf.* Smoryński, 1977, p. 858.

Pour la direction inverse, la situation est un peu plus compliquée. À première vue, l'argument informel suivant peut sembler faire l'affaire. Supposons que $\mathcal{F} \cup \mathcal{I}$ ne soit pas cohérente. Alors, toute formule de $\mathcal{L}_{\mathcal{F}\cup\mathcal{I}}$, le langage de $\mathcal{F} \cup \mathcal{I}$, est un théorème. En particulier, comme $\mathcal{L}_{\mathcal{F}}$ est inclus dans $\mathcal{L}_{\mathcal{F}\cup\mathcal{I}}$, on peut choisir une contradication $\bot$ prise dans $\mathcal{L}_{\mathcal{F}}$ et telle que $\mathcal{F} \cup \mathcal{I} \vdash \bot$. Mais alors, par conservativité, il s'ensuit que $\mathcal{F} \vdash \bot$. Ce qui contredit la consistance de $\mathcal{F}$. On aurait donc établi la cohérence de $\mathcal{F} \cup \mathcal{I}$ à partir de celle de $\mathcal{F}$ et d'une propriété de conservativité. Néanmoins, il n'est pas clair que ce raisonnement soit acceptable d'un point de vue finitiste. Le point problématique concerne la façon dont on peut exprimer la propriété de conservativité dans un cadre strictement finitiste. Si l'on veut dire que pour toute démonstration d'une formule finitiste φ dans $\mathcal{F} \cup \mathcal{I}$, il en existe une démonstration dans $\mathcal{F}$, on se heurte au fait que cet énoncé comporte une quantification exitentielle non-bornée et nous conduit aussitôt hors du cadre des mathématiques finitistes.[33] Il faut donc être attentifs à la façon dont on saisit la propriété de conservativité au sein de $\mathcal{F}$ et nous aurons besoin d'une hypothèse un peu plus forte. Donc, supposons la conservativité de $\mathcal{F} \cup \mathcal{I}$ sur $\mathcal{F}$ démontrée d'une manière acceptable d'un point de vue finitiste. Plus précisément, supposons que nous ayons une méthode admissible d'un point de vue finitiste qui permette de transformer toute preuve dans $\mathcal{F} \cup \mathcal{I}$ en une preuve dans $\mathcal{F}$. En d'autre terme nous disposons d'une fonction f primitive récursive telle que

$$\text{Pour tout énocé } \varphi \text{ du langage de } \mathcal{F},$$
$$\mathcal{F} \vdash \forall x(Pr_{\mathcal{F}\cup\mathcal{I}}(x, \overline{\varphi}) \to Pr_{\mathcal{F}}(f(x), \overline{\varphi}))$$ [34]

On peut alors raisonner de la manière suivante :

$$\mathcal{F} \vdash \forall x(Pr_{\mathcal{F}\cup\mathcal{I}}(x, \bot) \to Pr_{\mathcal{F}}(f(x), \bot))$$
$$\mathcal{F} \vdash \forall x(\neg Pr_{\mathcal{F}}(f(x), \bot) \to \neg Pr_{\mathcal{F}\cup\mathcal{I}}(x, \bot))$$
$$\mathcal{F} \vdash \forall x \neg Pr_{\mathcal{F}}(f(x), \bot) \to \forall x \neg Pr_{\mathcal{F}\cup\mathcal{I}}(x, \bot)$$
$$\mathcal{F} \vdash \forall x \neg Pr_{\mathcal{F}}(x, \bot) \to \forall x \neg Pr_{\mathcal{F}\cup\mathcal{I}}(x, \bot)$$
$$\textit{i.e. } \mathcal{F} \vdash Con(\mathcal{F}) \to Con(\mathcal{F} \cup \mathcal{I})$$

33. Ainsi, par exemple, l'énoncé $\exists x Pr_{\mathcal{F}\cup\mathcal{I}}(x, \overline{\varphi}) \to \exists x' Pr_{\mathcal{F}}(x', \overline{\varphi})$ n'est pas Π^0_1.

34. Où $Pr_{\mathcal{F}\cup\mathcal{I}}$ et $Pr_{\mathcal{F}}$ désignent respectivement un prédicat de prouvabilité dans $\mathcal{F} \cup \mathcal{I}$ et dans $\mathcal{F}$, et où $\overline{\varphi}$ est un nom (ou un code) désignant φ. L'énoncé $\forall x(Pr_{\mathcal{F}\cup\mathcal{I}}(x, \overline{\varphi}) \to Pr_{\mathcal{F}}(f(x), \overline{\varphi}))$ est Π^0_1 et peut alors se « lire » comme : si x est une preuve de φ dans $\mathcal{F} \cup \mathcal{I}$, alors $f(x)$ en est une preuve dans $\mathcal{F}$. Ce qui est une manière (forte, au sens où elle s'appuie sur f, une méthode de transformation *explicite* des preuves dans $\mathcal{F} \cup \mathcal{I}$ en des preuves dans $\mathcal{F}$) d'énoncer la conservativité syntaxique de $\mathcal{F} \cup \mathcal{I}$ sur $\mathcal{F}$.

On a donc réduit la cohérence de $\mathcal{F} \cup \mathcal{I}$ à celle de $\mathcal{F}$, et ce en n'utilisant que des moyens finitistes. Si, comme Hilbert semble l'avoir suggéré, on tient pour acquis dès le départ la cohérence des mathématiques finitistes, on peut en quelque sorte décharger l'antécédent de l'implication ci-dessus pour obtenir une démonstration finitiste de la cohérence de l'ensemble des mathématiques :

$$\mathcal{F} + Con(\mathcal{F}) \vdash Con(\mathcal{F} \cup \mathcal{I})$$

Ne pourrait-on pas faire mieux et exiger $\mathcal{F} \vdash Con(\mathcal{F} \cup \mathcal{I})$? Après tout, l'ensemble des moyens finitistes sont censés être formalisés par $\mathcal{F}$ *tout seul*, et non par $\mathcal{F} + Con(\mathcal{F})$. Bien sûr, venant après Gödel, nous pouvons douter que l'on puisse se passer de l'hypothèse $Con(\mathcal{F})$. Pour Hilbert, qui développait son Programme avant l'apparition des théorèmes d'incomplétude, la situation n'était sans doute pas aussi claire. Puisque d'un côté il affirme que la consistance des mathématiques est non problématique et ne saurait être réduite à quoi que ce soit d'autre, et que d'autre part, il considère que l'ensemble des mathématiques finitistes sont incluses dans $\mathcal{F}$, il est très probable qu'il ait considéré que $Con(\mathcal{F})$ était déjà présent dans $\mathcal{F}$ (ou trivialement démontrable à partir de $\mathcal{F}$). Autrement dit, il est très vraisemblable qu'Hilbert ait cru que $\mathcal{F} \vdash Con(\mathcal{F})$. Cette hypothèse est évidemment fausse pour toute théorie soumise aux théorèmes d'incomplétude. [35]

Au moment où Hilbert mettait au point son Programme dans les années 1920, de nombreux succès partiels laissaient à penser qu'il pourrait bientôt être mené à terme. Néanmoins, au tournant des années 1930, les théorèmes d'incomplétude vinrent mettre à mal cet optimisme. En exhibant un énoncé du langage de l'arithmétique équivalent à une formule Π^0_1 mais indémontrable dans un système [36] dont on pouvait penser qu'il regroupait toutes les mathématiques finitistes, le premier théorème d'incomplétude sonnait comme une réfutation du Programme de Conservativité. Quant au second théorème d'incomplétude, en montrant que toute théorie T contenant un minimum d'arithmétique ne pouvait prouver sa propre cohérence (*i.e.* $T \nvdash Con(T)$), il semblait contredire le Programme de Consistance. Lorsqu'ils furent démontrés les théorèmes d'incomplétude ont bien constitué un coup d'arrêt du projet hilbertien. Plus récemment cependant, cette interprétation a été remise en cause et l'on a assisté à un véritable renouveau des recherches s'inscrivant dans la ligne du Programme de Hilbert. [37]

35. En particulier, si on identifie les mathématiques finitistes à l'arithmétique primitive récursive, on sait que $\mathcal{PRA} \nvdash Con(\mathcal{PRA})$.

36. Le système P des *Principia*, qui est lui-même plus fort que l'arithmétique primitive récursive.

37. Une description exhaustive de ce renouveau des études hilbertiennes sortirait du cadre de ce travail.

3.1.3 Conservativité au service du nominalisme : Field (1980)

La défense Hilbertienne du finitisme n'est pas le seul exemple d'une argumentation d'inspiration « déflationniste », visant à montrer l' « absence de substance » [38] et le manque « pouvoir explicatif » de certains de nos concepts par un résultat de conservativité. Le programme nominaliste de Hartry Field en est une autre illustration célèbre, plus récente. Dans *Science Without Numbers* (1980), Hartry Field se propose de défendre une position originale en philosophie des mathématiques qu'on pourrait appeler nominalisme fictionnaliste. Field est nominaliste en ce qu'il refuse toute existence aux objets abstraits, en particulier aux objets mathématiques. Par conséquent, selon lui, les mathématiques —et Field entend par là *toutes* les mathématiques— dans la mesure où elles contiennent des références à des entitités abstraites, ne sont pas, à strictement parler, vraies. [39] Elles ne sont que des fictions commodes qui trouvent leur emploi dans la formulation de nos théories scientifiques. [40]

En rejetant tout engagement ontologique envers les objets abstraits, il est clair que Field s'oppose principalement, et radicalement, au platonisme. C'est d'ailleurs contre cette dernière philosophie des mathématiques que Field développe ses arguments. Le projet de Field ne consiste pas tant à argumenter directement en faveur du nominalisme qu'à priver la position adverse, à savoir le réalisme vis-à-vis des entités mathématiques, de ce qu'il considère comme le seul argument valable en faveur de l'existence d'objets mathématiques : un argument d'indispensabilité dû à Quine et à Putnam. [41] Sommai-

On pourra néanmoins consulter : Detlefsen, 1986, 1990, 1979, 2001 ; Raatikainen, 2003 ; Simpson, 1988, 2009.

38. ou encore, pour reprendre la terminologie de Hilbert lui-même : l'absence de contenu.

39.

> Envers cette part des mathématiques qui contient effectivement des références à (ou des quantifications sur) des entités abstraites —et ceci inclut virtuellement toutes les mathématiques habituelles— j'adopte une attitude fictionnaliste : c'est-à-dire que je ne vois aucune raison de regarder cette part des mathématiques comme vraie. (Field, 1980, p. 2)

40. Notons que, contrairement à d'autres variantes nominalistes en philosophie des mathématiques, et contrairement à Hilbert également, le nominalisme de Field n'est pas révisionniste concernant la signification des énoncés mathématiques. Les formules et les théorèmes de nos mathématiques conventionnelles ont bien le sens qu'ils semblent avoir, *i.e.* ils réfèrent bien à des objets abstraits dont certains sont infinis, tels que les nombres, les fonctions, les ensemble *etc.* Simplement, ces énoncés sont tous entièrement faux. Ce qui ne veut pas dire qu'ils ne sont pas utiles.

41.

> . . . le seul argument en faveur de la conception selon laquelle les mathématiques sont vraies, qui ne soit pas une pétition de principe [the only non-question begging argument *for* the view that mathematics consists of truths] (Field, 1980, p. 4).
>
> . . . il apparaît clairement qu'il n'y a qu'un seul et unique argument sérieux en faveur de

rement, l'argument dit de Quine/Putnam consiste à soutenir qu'un engagement ontologique envers les mathématiques découle du fait que les énoncés mathématiques font partie intégrante de nos meilleures théories physiques et que, par conséquent, la confirmation empirique de ces dernières vaut aussi confirmation de leur part mathématique. Cet argument est disséminé dans de nombreux ouvrages de Quine [42] et de Putnam, mais n'a que rarement été explicitement formulé ou entièrement développé par ces auteurs. [43] En voici une version précise, due à Mark Colyvan (2001, p.11) :

(1) Nous sommes tenus de nous engager ontologiquement envers *toutes* [44] les entités qui sont indispensables à nos meilleures théories scientifiques

(2) Les entités mathématiques sont indispensables à nos meilleures théories scientifiques

(C) Par conséquent, nous sommes tenus de nous engager ontologiquement envers les entités mathématiques

Chez Quine comme chez Putnam, cet argument s'inscrit dans un large cadre méthodologique qui sous-tend toute leurs philosophies. Le premier axe de ce cadre méthodologique est le naturalisme : le philosophe doit prendre acte du fait que les sciences sont notre méthode la plus sûre pour nous enquérir de la structure du réel. Il doit donc renoncer à toute velléité de philosophie première : il n'existe pas de point de vue extérieur et supérieur à celui des sciences pour juger la réalité. Pour les question concernant l'existence de telle ou telle entité, les sciences constituent donc l'ultime arbitre. Le second axe de ce cadre méthodologique est ce qu'on appelle habituellement le holisme de la confirmation. Nos théories scientifiques affrontent le tribunal de l'expérience « en bloc ». En cas de données empiriques récalcitrantes, la révision de nos théories peut atteindre jusqu'aux parties les plus centrales de notre schème conceptuel, telles que la logique ou les mathématiques. À l'inverse lorsque nos théories sont confirmées par l'expérience et que nous sommes justifiés à les croire, cette confirmation et cette justification portent aussi sur les portions de nos théories les plus éloignées, en apparence, de l'expérience. Ainsi, la

> l'existence des entités mathématiques, et qu'il s'agit de l'argument quinien selon lequel nous avons besoin de postuler de telles entités pour pouvoir effectuer des inférences ordinaires à propos du monde phyisque et pour pouvoir faire des sciences. (Field, 1980, p. 5)

42. Du moins après qu'il eut renoncé au nominalisme qu'il défendait encore aux côtés de Goodman en 1947. *Cf.* Goodman et Quine, 1947.

43. À notre connaissance, la formulation la plus complète de cet argument d'inspensabilité par l'un ses auteurs d'origine se trouve dans Putnam, 1971.

44. *Toutes* : c'est-à-dire y compris envers les entités mathématiques qui s'avéreraient indispensables à nos théories scientifiques.

confirmation empirique de nos théories physiques sanctionne aussi la logique et les mathématiques qui y sont employées. Ces deux hypothèses méthodologiques fondamentales fournissent une justification à la première prémisse du raisonnement ci-dessus. Le naturalisme nous conduit à ne devoir accepter que les entités postulées par les sciences, tandis que le holisme de la confirmation nous contraint à traiter sur un pied d'égalité ontologique toutes les entités apparaissant dans nos meilleures théories scientifiques, que ce soient des objets macroscopiques directement observables ou des objets abstraits comme ceux des mathématiques. [45]

Pour combattre cette conception, Field développe un Programme nominaliste. Il faut souligner que, dans l'ensemble, Field embrasse l'arrière-plan méthodologique quinien que nous venons de décrire, et que par conséquent il accepte la première prémisse du raisonnement ci-dessus. Son Programme vise donc à réfuter la seconde prémisse. Il s'agit pour lui de montrer que les mathématiques ne sont pas indispensables à nos meilleures théories scientifiques. Pour ce faire, Field adopte une stratégie en deux temps. Tout d'abord, il faut montrer qu'on peut construire une version « nominaliste » de nos principales théories scientifiques, c'est-à-dire en donner une formulation dans laquelle n'apparaissent pas de référence ou de quantification portant sur des entités abstraites. Dans son ouvrage, [46] Field illustre comment une telle opération est possible en développant en détail une version nominalisée de la théorie newtonienne classique de la gravitation. [47] Cet exemple

45. On trouvera une discussion très éclairante des liens entre, d'une part, le naturalisme et le holisme de la confirmation quiniens, et, d'autre part, l'argument d'inspensabilité dans Colyvan, 2001. Pour être tout à fait exhaustif et précis, il faut ajouter aux deux principes méthodologiques mentionnés, un critère d'engagement ontologique des théories. À (l'existence de) quoi nous engage telle théorie que nous avons acceptée ? Quelles entités ou quels objets existent d'après cette théorie ? Selon la célèbrissime formule de Quine : « être c'est être la valeur d'une variable » (Quine, 1948), les entités que nous sommes tenus d'accepter lorsque nous adoptons une théorie sont celles auxquelles les termes de cette théorie réfèrent ou sur lesquelles portent ses quantifications. Les théories physiques habituelles (c'est-à-dire avant qu'elles aient été soumises à l'éventuelle nominalisation proposée par Field, si tant est qu'une telle nominalisation soit possible —voir les paragraphes suivants) contiennent des termes mathématiques et des quantifications portant sur des objets mathématiques. À première vue, elles semblent donc s'accompagner d'un engagement ontologique vis-à-vis d'entités mathématiques abstraites.

46. Field, 1980

47. Les détails techniques de cette nominalisation importent peu ici. Disons simplement que Field généralise un théorème de représentation dû à Hilbert pour, d'une part, remplacer les quantifications sur des nombres réels par des quantifications sur des points de l'espace, et, d'autre part, éviter la référence à des notions mathématique comme la distance entre deux points en s'appuyant sur des prédicats comparatifs portant sur les points tels « x est situé entre y et z » ou « xy est congruent à zw ». Les quantifications sur les nombres ou sur les ensembles de nombres sont ainsi remplacées dans la physique nominalisée par des quantifications sur des points de l'espace-temps et sur des « régions » de points de l'espace-temps. Les grandeurs quantitatives de la physique telles que « x a une masse de ... » sont elles remplacées, à l'instar de la notion de distance, par des référence à des relations comparatives entre points de l'espace-temps telles que « x a un plus grand potentiel gravitationnel que y ». *Cf.* Field, 1980 ou

est proposé par Field à titre de paradigme, comme première brique d'un programme général plus large ambitionnant de nominaliser toutes les théories scientifiques. Dans un second temps, Field doit expliquer pour quelle raison les sciences ont eu recours au mathématiques. Comment celles-ci peuvent être utiles aux scientifiques alors même, que selon Field, elles sont entièrement fausses ? C'est ici qu'intervient la notion de conservativité. Pour Field les mathématiques n'ont pas besoin d'être vraies pour être utiles à la science. [48] Seul compte le fait qu'elles ne nous induisent pas en erreur. Lorsqu'on adjoint à un ensemble d'assertions formulées dans un langage purement nominaliste, un corps d'énoncés mathématiques, on ne doit pas obtenir de nouvelles conséquences exprimables dans le langage nominaliste. Autrement dit, l'ajout des mathématiques à une théorie physique nominalisée doit résulter en une extension conservative. Cette conservativité permet, à l'instar du Programme de Hilbert, d'envisager un usage purement instrumental des mathématiques. [49] Les fictions mathématiques peuvent être utiles pour raccourcir nos calculs ou faciliter nos inférences. Field illustre d'ailleurs comment elles simplifient grandement nos raisonnements. [50] Mais *in fine*, seules comptent les assertions nominalistes de nos théories puisqu'elles seules portent sur des objets existant réellement et qu'elles seules sont susceptibles d'être vraies. [51] Les mathématiques, bien qu'utiles sur un plan pratique, sont néanmoins « dispensables » en principe. Et voici l'argument de Quine-Putnam réfuté.

Dans *Science without Numbers* Field reste assez vague quant au type de conservativité qu'il entend employer au sein de son Programme, tout comme sur le cadre logique dans lequel le nominalisation de nos théories physiques est censée avoir lieu. Au long

pour un résumé simplifié Colyvan, 2001, chapitre 4. Au passage, on voit que Field admet dans la classe des objets concrets, les points de l'espace de la physique de Newton ainsi que les régions de points de cet espace. En d'autres termes, ces *geometricalia* sont considérés par Field comme existants et concrets, ce qui n'est pas du goût de tous les nominalistes (voir, par exemple : Resnik, 1985).

48.

> « Une thèse centrale de *Science without Numbers* est que les mathématiques n'ont pas à être vraies pour être utiles [A central claim of *Science without Numbers* is that mathematics does not need to be true to be good] » Field, 1985, p. 125

49. Cette analogie tracée entre le programme de Field et le finitisme hilbertien est assez naturelle. Une telle comparaison est d'ailleurs explicitement proposée par exemple dans Shapiro, 1983, ainsi que dans Urquhart, 1990.

50. Au demeurant, la version nominalisée de la physique newtonienne est d'une formulation pour le moins assez contournée et difficilement maniable.

51. Ou faussses si, comme c'est le cas pour la physique newtonienne, nos théories sont erronées. Comme nous l'avons déjà signalé, les assertions purement mathématiques sont, quant à elles, toutes uniformément fausses selon Field.

des chapitres, il propose plusieurs versions techniques de son projet : certaines dans une logique du premier ordre, d'autres articulées sur des logiques plus fortes. De même, certains passages semblent suggérer que les mathématiques doivent être déductivement conservatives sur une physique nominaliste, tandis que d'autres paraissent privilégier la conservativité sémantique. [52] Cependant, lorsqu'il applique la nominalisation à l'exemple paradigmatique de la théorie newtonienne de la gravitation, Field se place clairement dans une logique de second ordre. [53] La situation semble donc être la suivante : une théorie nominaliste de la physique (newtonienne) $\mathcal{PN}$ ne contenant pas d'objets mathématiques est développée ; elle doit être telle que, lorsqu'on lui adjoint une théorie mathématique « raisonnable » $\mathcal{M}$, alors $\mathcal{PN} \cup \mathcal{M}$ est une extension conservative de $\mathcal{PN}$. [54]

Mais cette caractérisation est quelque peu ambiguë : faut-il entendre ici que la théorie étendue $\mathcal{PN} \cup \mathcal{M}$ est sémantiquement conservative sur $\mathcal{PN}$ ou bien qu'elle est déductivement conservative ? Un intéressant phénomène se produit ici, comme l'a montré Stewart Shapiro, 1983. Dans la construction proposée par Field, $\mathcal{PN} \cup \mathcal{M}$ est bien sémantiquement conservative sur $\mathcal{PN}$. Tout modèle de $\mathcal{PN}$ peut se prolonger en un modèle de $\mathcal{PN} \cup \mathcal{M}$ et tout énoncé nominaliste qui est conséquence sémantique de notre théorie nominaliste augmentée de mathématiques est déjà conséquence sémantique de la seule théorie nominaliste. [55] Cependant Shapiro, 1983 montre que la théorie étendue n'est pas déductivement conservative, ce qui affaiblit considérablement la portée du résultat obtenu par Field. Plus spécifiquement, même si l'ajout de $\mathcal{M}$ ne modifie pas la classe des énoncés qui sont conséquences sémantiques de notre théorie, elle augmente nos capacités de démonstration : il existe des énoncés physiques qui ne nous sont pas accessibles sans détour par $\mathcal{PN} \cup \mathcal{M}$, au sens où on ne pourrait les prouver, fut-ce de manière fasti-

52. Sur les diverses formulations techniques possibles d'une propriété de conservativité des mathématiques et leurs liens avec le projet fieldien voir Urquhart, 1990.

53. Il emploie en effet non seulement une quantification sur les points de l'espace-temps (au moyen de variables d'individus) mais aussi sur les régions de l'espace-temps, considérés comme des sommes méréologiques de points (au moyen de variables d'ordre supérieur).

54. Voir (Field, 1980, chapitres 6 à 8, p. 47–92). Nous simplifions quelque peu. En réalité, lorsqu'on étend $\mathcal{PN}$ en lui adjoignant une théorie mathématique $\mathcal{M}$, on étend aussi le domaine de quantification des formules de $\mathcal{PN}$. En outre, si $\mathcal{PN}$ exclut l'existence d'entités abstraites alors $\mathcal{PN} \cup \mathcal{M}$ sera inconsistante. Pour contourner ce problème technique sans importance ici, Field introduit un prédicat $M(x)$ défini comme « x est une entité mathématique » et traduit tout énoncé φ dont le vocabulaire non logique est cantonné au seul langage de $\mathcal{PN}$ en une assertion purement nominale φ^* en relativisant tous ses quantificateurs à l'aide de la formule « $\neg M(x)$ ». C'est alors la conservativité des mathématiques sur $(\mathcal{PN})^*$ qui sera cruciale.

55. Ceci est dû au fait que la théorie du second ordre $\mathcal{PN}$ est catégorique.

dieuse, inélégante ou arbitrairement longue, sans employer la théorie mathématique $\mathcal{M}$. $\mathcal{M}$ peut donc être au moins considérée comme « épistémiquement indispensable ». [56] Nous ne poursuivrons pas ici plus avant la discussion de cette question. Notre rapide exposé du Programme de Field avait seulement pour but d'illustrer, comment, après l'entreprise hilbertienne, la notion de conservativité a pu être à nouveau mise au service d'un argument d'inspiration généralement « déflationniste » [57] vis-à-vis, cette fois, de toutes les entités mathématiques. Néanmoins, il y a certains points que nous voudrions mettre en exergue, car ils ont leur importance pour la discussion à venir concernant la conception déflationniste de la vérité.

Tout d'abord, il y a, nous semble-t-il, un lien important tracé dans le nominalisme de Field entre engagement ontologique et propriété de (non)-conservativité. La conservativité des mathématiques sur notre théorie physique est censée justifier une attitude fictionnaliste vis-à-vis de leurs objets et un emploi instrumental des outils théoriques qu'elles fournissent : les entités mathématiques n'existent pas, les énoncés qui y réfèrent sont faux, mais les théories mathématiques n'ont pas besoin d'être vraies pour être utiles, dès lors qu'elles sont conservatives. Mais, à l'inverse, les objets ou les concepts dont l'ajout se révèle non-conservatif ne doivent-ils pas être considéré comme ayant une existence réelle et un contenu substantiel ? Sans doute. C'est que l'anti-réalisme fictionnaliste de Field se limite aux mathématiques. Field est en revanche parfaitement réaliste en ce qui concerne les entités théoriques inobservables qui apparaissent dans nos meilleures théories physiques, tout comme il croit au contenu « substantiel » des propriétés qui leur sont attribuées. Pourtant, comme Field le reconnait lui-même, on pourrait pousser l'entreprise de reformulation de la physique plus loin, jusqu'à obtenir une théorie physique $\mathcal{PO}$ [58], où seuls apparaitraient des termes référant à des entités macroscopiques *directement observables*. Mais, selon Field, à la différence du cas des mathématiques, lorsqu'on étendra $\mathcal{PO}$ en lui joignant une théorie $\mathcal{T}$ formalisant le contenu de notre physique portant sur les entités inobservables, on obtiendra une théorie étendue $\mathcal{PO} \cup \mathcal{T}$

56. *Cf.* Shapiro, 1983 et la réponse de Field (Field, 1985). Notons que dans ce texte de 1985, Field reconnait qu'il n'a pas été assez clair dans sa monographie mais affirme que c'est bien la conservativité sémantique qui importe pour son projet. Quelques années plus tard (*cf.* Field, 1990), suite aux débats que son projet aura suscités, il sera néanmoins tout proche de renoncer à la logique du second ordre qu'il défendait jusque là pour adopter finalement une logique du premier ordre. Dans ce cas, ainsi que nous l'avons déjà expliqué, la distinction entre conservativité déductive et conservativité sémantique n'a plus grande pertinence.

57. Au sens, un peu vague il est vrai, où il s'agit de montrer que certains des concepts qui apparaissent dans nos théories ne sont pas « substantiels » et ne jouent pas de rôle explicatif réel.

58. Plus stricte ou plus ontologiquement économe encore que la physique nominalisée $\mathcal{PN}$.

qui ne sera non-conservative sur $\mathcal{PO}$. Cette différence est cruciale. La non-conservativité des entités théoriques, autres que mathématiques, est la manifestation du rôle explicatif qu'elles jouent dans nos entreprises scientifiques. C'est ce qui les rend indispensables à nos meilleures théories du monde et, dès lors que l'on adhère au principe méthodologique quinien du naturalisme, c'est également ce qui doit nous conduire à accepter leur existence et à considérer les termes qui y réfèrent comme dôté d'un contenu « substantiel ». Ce point est bien illustré par le passage suivant extrait du premier chapitre de *Science without Numbers*.[59] Field y compare le cas des entités théoriques inobservables (à travers l'exemple des particules subatomiques) et celui des mathématiques, après avoir remarqué que pour pouvoir les employer au sein de nos inférences scientifiques, il est nécessaire d'introduire des lois-ponts reliant ces entités aux objets observables, :

> « *Mais il existe une différence fondamentale entre ces deux cas, et cette différence réside dans la nature des lois-ponts.* Dans le cas des particules subatomiques, la théorie $\mathcal{T}$, à présent interprétée de manière à inclure les lois-ponts (ainsi que, peut-être, quelques hypothèses concernant les conditions initiales), peut être appliquée à un ensemble de prémisses portant sur les observables d'une manière à engendrer des assertions authentiquement nouvelles à propos des observables, assertions qui ne seraient pas dérivables sans l'aide de $\mathcal{T}$. Mais, dans le cas des mathématiques, la situation est tout à fait différente : ici, si nous prenons une théorie mathématique qui comprend les lois-ponts (*i.e.* qui comprend des énoncés assertant l'existence de fonctions reliant les objets physiques à certains objets abstraits "purs", et peut-être notamment des assertions obtenues au moyen d'un principe de compréhension qui emploie en même temps et du vocabulaire physique, et du vocabulaire mathématique), alors les mathématiques sont applicables au monde, *i.e.* elles sont utiles en ce qu'elles nous permettent de tirer des conclusions nominalistiquement formulables à partir de prémisses nominalistiquement formulables ; *mais là, contrairement au cas de la physique, les conclusions auxquelles nous parvenons par cet intermédiaire ne sont pas authentiquement nouvelles, elles sont déjà dérivables d'une façon plus ardue à partir des prémisses, sans recours aux entités mathématiques.* » (Field, 1980, p. 10–11, italiques de l'au-

59. Justement intitulé : *Pourquoi l'utilité des mathématiques est différente de l'utilité des entités théoriques.*

teur)

Ainsi, ce qui distingue les fictions mathématiques d'autres entités théoriques réelles qui s'affichent dans nos théories, c'est leur caractère conservatif. *A contrario*, la non-conservativité de certaines entités théoriques, comme les particules subatomiques, justifie qu'on prennent au sérieux leur existence et qu'on leur attribue un contenu « substantiel ». C'est un point dont nous devrons nous souvenir quand nous examinerons la question de la vérité déflationniste.

Un second élément que nous voulons signaler à propos du nominalisme de Field et des débats qu'il a provoqués concerne la critique de Shapiro (1983) que nous avons déjà évoquée. Pour montrer la non-conservativité déductive des mathématiques sur la physique newtonienne nominalisée, [60] Shapiro a recours à une construction technique tout à fait intéressante. L'argument de Shapiro s'appuie en fait sur les théorèmes d'incomplétude de Gödel. Pourtant ces théorèmes sont la plupart du temps énoncés dans le contexte des théories mathématiques, en particulier arithmétiques. Néanmoins, la stratégie de Shapiro consiste à reconstruire les théorèmes de Gödel en se limitant aux seuls moyens de la physique newtonienne nominalisée. Il montre que la théorie $\mathcal{PN}$ contient des ressources suffisantes pour que l'on puisse définir à l'intérieur de cette théorie une région R de l'espace constituée d'une suite discrète de points, contenant un premier élément dénoté p, et telle que la distance entre chaque point est uniforme. Cette région R est un « analogue physique » des entiers naturels. [61] On peut alors définir sur cette région R des « analogues physiques » des principales opérations de l'arithmétique et ensuite coder la syntaxe du langage de $\mathcal{PN}$ dans cette structure (R, p) composée de points « physiques ». Ainsi, l'arithmétisation de la syntaxe est ici remplacée par une géométrisation de la syntaxe qui a lieu dans l'espace physique concret de $\mathcal{PN}$. Une fois cette géométrisation réalisée, on peut imiter la démarche gödelienne et exhiber un énoncé du langage de $\mathcal{PN}$ qui affirme que la théorie $\mathcal{PN}$ est cohérente. Cet énoncé $Con_{\mathcal{PN}}(R, p)$ a aussi, en un sens, un contenu physique : il énonce certaines propriétés reliant les points de (R, p). La construction proposée par Shapiro est cependant telle que les conditions d'application du second théorème de Gödel sont remplies. Et on a donc $\mathcal{PN} \nvdash Con_{\mathcal{PN}}(R, p)$. En revanche, la théorie étendue $\mathcal{PN} \cup \mathcal{M}$ contient assez de théorie des ensembles pour

60. c'est-à-dire, reprenant la notation que nous avons introduite un peu plus haut, la non-conservativité de $\mathcal{PN} \cup \mathcal{M}$ sur $\mathcal{PN}$.

61. Elle est une ω-séquence formée de points de l'espace qui, selon les propres critères de Field, sont des objets concrets existant réellement.

prouver $Con_{\mathcal{PN}}(R, p)$.[62] Autrement dit, $\mathcal{PN} \cup \mathcal{M} \vdash Con_{\mathcal{PN}}(R, p)$, ce qui montre la non-conservativité déductive des mathématiques sur la physique. Là encore, il sera utile de garder en mémoire cette construction de Shapiro lors de l'examen de certains arguments concernant la vérité déflationniste.

Il est frappant de constater que l'on retrouve exactement les mêmes « ingrédients » techniques[63] dans la discussion de l'argument de la conservativité à propos du déflationnisme en matière de vérité.

3.1.4 Vérité déflationniste et conservativité

Comme l'illustrent les deux exemples historiques que nous avons rappelés ci-dessus, la notion de conservativité a souvent été employée dans des arguments visant à « dégonfler » certains concepts apparaissant dans nos discours théoriques sur le monde.

Etant donné les thèses défendues par les déflationnistes concernant la nature de la vérité, il n'est guère étonnant que la notion de conservativité ait été invoquée pour capturer l'absence de « substantialité » que les déflationnistes attribuent au prédicat « vrai ». Dans (Shapiro, 1998b), Stewart Shapiro affirme ainsi que si le prédicat de vérité n'est qu'un « outil logico-syntaxique de décitation » , utile et même indispensable pour pouvoir exprimer certaines généralisations mais dépourvu de contenu explicatif propre, s'il ne dénote pas une propriété réelle ou si la notion de vérité n'est pas métaphysiquement substantielle, alors les théories déflationnistes de la vérité doivent satisfaire une contrainte de conservativité. Ketland a lui aussi, et de manière indépendante, proposé une caractérisation similaire dans (Ketland, 1999). Voici l'extrait dans lequel Shapiro justifie la contrainte de conservativité :

> Je suggère que la conservativité sous une forme ou sous une autre[64] est essentielle pour le déflationnisme. Supposons, par exemple que Karl adhère de façon justifiée à une théorie $\mathcal{B}$ dans un langage qui ne permet pas d'ex-

62. Nous simplifions quelque peu la construction de Shapiro. En fait, le raisonnement de Shapiro s'appuie aussi sur l'existence, *dans* $\mathcal{PN} \cup \mathcal{M}$ d'un homomorphisme de réprésentation reliant les points de l'espace-temps de $\mathcal{PN}$ et $\mathbb{R}^4$, ce qui permet d'établir *dans* $\mathcal{PN} \cup \mathcal{M}$ un isomorphisme entre (R, p) et $(\omega, 0)$. Pour plus de détails, voir Shapiro, 1983, p. 526–527.

63. Ainsi que, pour partie, les mêmes protagonistes.

64. Shapiro ne précise pas dans ce passage à quel type de conservativité (déductive ou sémantique) il fait référence. Cependant, dans le cas de théories formalisées —ou formalisables— dans une logique du premier ordre (et c'est ainsi que se présentent en premier abord la théorie minimale ou l'axiomatisation par l'ensemble des T-équivalences décitationnelles), on sait que les deux notions coïncident. Cette ambiguité n'est par conséquent pas gênante.

> primer la vérité. Il ajoute au langage un prédicat de vérité et étend $\mathcal{B}$ en une théorie $\mathcal{B}'$ en n'usant que d'axiomes essentiels à la vérité. Supposons que $\mathcal{B}'$ n'est pas conservative sur $\mathcal{B}$. Alors, il existe un énoncé Φ du langage d'origine (en sorte que Φ ne contient pas le prédicat de vérité) qui est conséquence [65] de $\mathcal{B}'$ mais qui n'est pas conséquence de $\mathcal{B}$. C'est-à-dire qu'il est logiquement possible que les axiomes de $\mathcal{B}$ soient vrais et que cependant Φ soit fausse, mais il est logiquement impossible que les axiomes de $\mathcal{B}'$ soient vrais et que Φ soit fausse. Ceci mine le thème déflationniste central selon lequel la vérité est non substantielle. Avant que Karl n'adhère à $\mathcal{B}'$, $\neg\Phi$ était possible. Le passage de $\mathcal{B}$ à $\mathcal{B}'$ a *ajouté* un contenu sémantique suffisant pour exclure la fausseté de Φ. Mais par hypothèse, seuls ont été ajoutés dans $\mathcal{B}'$ des principes essentiels à la vérité. Donc, ces principes ont un contenu sémantique substantiel. (Shapiro, 1998b, p. 498, nous traduisons, italiques de l'auteur)

Le passage ci-dessus montre clairement que Shapiro relie substantialité et non-conservativité. En cela, il s'inscrit, tout comme Ketland, dans une perspective méthodologique héritée de celle qui fut mise en oeuvre d'abord par Hilbert puis par Field. A première vue, ce mouvement semble assez justifié.

Comme nous l'avons déjà longuement rappelé, les déflationnistes refusent à la vérité tout rôle explicatif et considèrent que le prédicat « vrai » ne désigne pas un propriété réelle ou « substantielle ». Mais comment donner un sens plus précis à ces thèses ? La notion de rôle ou de pouvoir explicatif est, reconnaissons-le, assez vague. Et, les débats restent ouverts en philosophie concernant la bonne manière d'expliquer la notion... d'explication. De même, la question de savoir si un terme de notre langage désigne ou non une propriété « substantielle », dotée d'un contenu « réel » n'est guère plus claire. [66] Néanmoins, il semble assez sensé de supposer que la capacité explicative d'une notion réside précisément dans le fait que l'on peut dériver ou prouver plus de choses à l'intérieur de nos théories une fois qu'on l'a adoptée. Si, de plus, nous adoptons un cadre méthodologique *inspiré* du naturalisme de Quine, alors nous sommes engagés à croire

65. Là aussi, Shapiro ne précise pas s'il entend parler de conséquence sémantique ou de conséquence déductive. Mais, dès lors que l'on s'intéresse à des théories du premier ordre, une remarque similaire à celle de la note précédent s'applique.

66. Si cette distinction peut sembler claire dans certains cas triviaux où l'intuition préthéorique et le bon sens peuvent suffire — *e.g.* « être un schtroumph » *versus* « être composé de fer »—, elle le sera beaucoup moins dans les cas litigieux. Or c'est justement ce qui nous occupe avec le déflationnisme en matière de vérité.

en l'existence réelle des entités et des propriétés qui sont indispensables à nos meilleures théories scientifiques. Pouvoir explicatif et « substantialité » se révèlent comme les deux faces d'une même médaille : seront acceptés comme pourvues de « substance » les entités qui apparaissent de manière indispensable dans nos meilleures théories du monde, c'est-à-dire celles qui jouent un rôle siginificatif dans nos explications scientifiques.

C'est en tout cas, nous semble-t-il, une conception de ce genre qui sous-tend l'appel à la notion de conservativité dans les deux exemples historiques qui précèdent. Chez Field, le lien entre « substantialité », pouvoir explicatif et non-conservativité est particulièrement clair dans le passage comparant les mérites des entités inobservables et des objets mathématiques que nous avons cité plus haut. Et, s'il serait en revanche évidemment anachronique d'invoquer à propos de Hilbert le naturalisme tel qu'il fut défendu par Quine et ses successeurs, reste que Hilbert trace bien un lien entre un soupçon de nature *ontologique* quant à l'existence des objets transfinis et le problème de leur emploi au sein de nos pratiques inférentielles. Par la conservativité il entend bien montrer que les méthodes et les concepts des mathématiques idéales sont en principes superflus, évitables ou, pour parler comme Quine, « dispensables ». Sur ce point, on peut donc considérer que le finitisme de Hilbert a en quelque sorte anticipé certains des traits méthodologiques du naturalisme contemporain. De manière plus ou moins explicite, Shapiro et Ketland reprennent ce cadre d'analyse et l'adaptent au cas du déflationnisme.

Il est vrai que les contextes philosophiques du finitisme hilbertien, du nomnalisme fictionnaliste à la Field et du déflationnisme en matière de vérité sont quelque peu différents. Selon Shapiro et Ketland, on peut néanmoins y appliquer une approche méthodologique similaire. A chaque fois, une thèse philosophique concernant la nature (ou l'absence de nature) de certaines entités est traduite par, ou débouche sur, une propriété formelle de conservativité. Ceci permet à la fois de donner un sens technique précis à la thèse philosophique, tout en la mettant, si l'on peut dire, à l'épreuve. La propriété de conservativité apparaît comme un test ou une expérience cruciale, qui selon qu'elle sera ou non satisfaite viendra confirmer ou infirmer l'hypothèse philosophique de départ. La propriété de conservativité, dès lors qu'elle est vérifiée, est censée permettre deux choses : d'une part, éviter tout engagement ontologique envers les notions dont l'ajout produit une extension conservative, et, ce faisant, leur dénier tout pouvoir explicatif au sein de nos théories du monde. Et, d'autre part, expliquer et justifier l'usage purement instrumental ou uniquement expressif qu'on peut faire de ces entités dans notre pratique

habituelle de théorisation, en arguant du fait que la conservativité montre que cet emploi est « sans danger », alors même qu'il est commode sur un plan pratique.

Le passage par les théories formalisées et le recours à la notion de conservativité est donc censé permettre de donner une caractérisation rigoureuse des allégations de « non substantialité » et de « nature purement décitationnelle mais non explicative » formulées par les déflationniste à l'égard de la vérité. Voici une manière un peu plus précise de formuler cette première contrainte :

Contrainte de Conservativité. Supposons ainsi que l'on parte une théorie de base T, exprimée dans un langage formel $\mathcal{L}$. $\mathcal{L}$ ne contient pas de prédicat de vérité ni de vocabulaire sémantique. On supposera néanmoins que T est suffisamment riche [67] pour pouvoir représenter sa propre syntaxe. Supposons à présent qu'on étende $\mathcal{L}$ par un prédicat « Vr » et qu'on ajoute à T une théorie formelle de la vérité $V(T)$ (exprimée dans $\mathcal{L}' \supseteq \mathcal{L} \cup \{Vr\}$), on doit alors obtenir une théorie $T \cup V(T)$ conservative sur T.

La demande de conservativité de la théorie de la vérité est essentielle pour le déflationniste : si les axiomes de vérité $V(T)$ ne donnent pas une extension conservative sur la théorie initiale T, alors il s'ensuit qu'on peut *démontrer* à l'aide du prédicat de vérité, et non pas simplement *exprimer*, des énoncés formulés dans le langage de T qu'on ne pouvait pas prouver à partir de T seule. Dans ce cas, la notion de vérité augmente nos capacités de déduction, y compris concernant les énoncés de $\mathcal{L}$, c'est-à-dire des énoncés dont le contenu ne concerne pas explicitement la vérité. Par exemple, si notre théorie de base T concerne les pommes et axiomatise leurs propriétés, le fait d'ajouter à T une théorie de la vérité *non*-conservative $V(T)$ nous permettra de démontrer de nouveaux faits à propos...des pommes. Or, si la vérité était un simple outil logico-syntaxique ne dénotant pas une réelle propriété, on ne voit pas comment elle pourrait interférer avec des faits concernant les pommes. Dès lors que le prédicat de vérité n'est pas conservatif, il a bien une capacité explicative forte et un usage théorique fécond. Et, il semble clair que la notion de vérité a bien un contenu substantiel, ce qui contredit la nature déflationniste de la vérité. [68] Donc, le déflationniste paraît contraint de fournir une théorie de la vérité qui soit conservative. Mais, il semble que cette exigence de conservativité

67. Par exemple en supposant que T contient un minimum d'arithmétique.

68. Comme le remarque Shapiro Shapiro, 1998b, p. 498, l'argument ci-dessus suppose que l'on relie la notion informelle de « substantialité métaphysique » aux notions de contenu sémantique et de conséquence logique : une notion sera considérée comme « substantielle » si son introduction nous permet de prouver plus de choses, y compris des choses dans lesquelles cette notion n'est pas directement impliquée.

ne soit pas compatible avec certains emplois du prédicat de vérité dans des explications mathématiques.

3.2 Réflexivité et arguments sémantiques

Nous en venons maintenant à la seconde étape de l'argument de Shapiro et Ketland. Comme nous l'avons déjà expliqué, l'idée de ces auteurs est de montrer qu'il existe, outre la contrainte de conservativité,[69] une autre contrainte d'adéquation s'imposant aux théories de la vérité mais qui, dans certaines circonstances, est incompatible avec la conservativité. Leur raisonnement s'appuie sur l'existence de certains arguments sémantiques et sur l'examen des modes de justification de ce qu'on appelle en logique des principes de réflexion. Notons que si les principes de réflexions peuvent s'ajouter à n'importe quelle théorie, le raisonnement de Shapiro et Ketland prend une acuité particulière lorsqu'on l'examine à la lumière des phénomènes d'incomplétude gödeliens.

3.2.1 Principes de réflexion

Considérez l'énoncé suivant :

« Tous les théorèmes de l'arithmétique de Peano sont vrais ».

Cet énoncé exprime la correction[70] d'une certaine théorie arithmétique : une théorie est correcte si ce qu'elle démontre est vrai. L'énoncé nous permet d'exprimer notre confiance dans la fiabilité de nos méthodes de preuves en théorie des nombres. De manière plus générale, lorsqu'on considère une théorie formelle T, l'énoncé « tous les théorèmes de T sont vrais » est ce qu'on appelle en logique un principe de réflexion (pour T). Pour peu que l'on se place dans un métalangage et une métathéorie pour T qui renferment des ressources linguistiques suffisantes, à savoir, des noms pour chacun des énoncés du langage de T (ou encore suffisamment de ressources pour pouvoir formaliser la syntaxe du langage de T), un prédicat (ou une formule) exprimant la propriété « être (le nom d'un énoncé qui est) un théorème de T » et un prédicat de vérité pour le langage[71] de T, ce principe de réflexion peut s'écrire sous la forme :

69. Et la contrainte minimale donnée par la CONVENTION **T** de Tarski sur laquelle toutes les parties s'accordent.

70. En anglais on parle de la *soudness* d'une théorie.

71. Rappelons que le prédicat de vérité est un prédicat typé et qu'il ne s'applique qu'aux énoncés de langage de T c'est-à-dire uniquement aux énoncés du langage-*objet*. Sont donc exclus les formules dans lesquelles le prédicat s'applique à un énoncé contenant déjà le prédicat « vrai ».

$$(Ref) \ : \ \forall x(Thm_T(x) \rightarrow Vr(x))$$

Lorsque nous adoptons une théorie parce que nous la considérons comme vraie, il semble que nous soyons aussi tenus d'en accepter tous les théorèmes, voire que nous soyons tenus d'accepter l'énoncé général affirmant que tous les théorèmes de la théorie sont vrais. Ainsi, les principes de réflexion apparaissent comme une extension naturelle de la théorie de base T. Cependant, dans bien des cas la théorie étendue, disons $T \cup \forall x(Thm_T(x) \rightarrow Vr(x))$ est plus forte d'un point de vue logique que la théorie de base T. Autrement dit, le principe de réflexion n'est pas conséquence logique de T et la théorie étendue permet de prouver plus de choses que la théorie d'origine.[72] Dès lors se pose la question de la justification des principes de réflexion. Une fois acceptée T, sommes-nous réellement justifiés à accepter (Ref) ? Et si tel est le cas faut-il simplement prendre cet énoncé comme un nouvel axiome, ou devons-nous en fournir une justification, peut-être sous la forme d'une preuve ?

3.2.2 Arguments sémantiques et contrainte de réflexivité

Selon Shapiro et Ketland, il existe un manière naturelle de justifier l'adoption de principes de réflexion. Cette justification consiste à donner une preuve sémantique de (Ref) dans une extension aléthique[73] suffisamment forte de T. Selon eux, munis d'une théorie de base T et d'une théorie de la vérité pour le langage de T, nous devrions être en mesure de mener le raisonnement suivant :

Preuve sémantique du principe de réflexion.

1. Tous les axiomes de T sont vrais
2. Les règles d'inférence de T préservent la vérité
3. *Donc*, tous les théorèmes de T sont vrais.

Ainsi, le principe de réflexion pour T se voit démontré dès lors que, d'une part, notre théorie de la vérité est assez forte pour établir (entendez prouver) (1.) que tous les axiomes de T sont vrais et (2.) que les règles d'inférence préservent la vérité, et d'autre part qu'une fois établies ces deux prémisses nous sommes en mesure d'en conclure que

72. Un sens plus précis sera donné à ces affirmations dans ce qui suit. Le point crucial est évidemment que, en général, $T \cup \forall x(Thm_T(x) \rightarrow Vr(x))$ n'est pas conservative sur T.

73. par extension aléthique nous désignons une théorie de la vérité étendant notre théorie de base T au moyen d'axiomes pour la vérité.

(3.) tous les théorèmes de T sont vrais—ce qui, nous allons le voir, nécessite d'effectuer un raisonnement inductif sur la longueur de preuves et d'employer une récurrence portant sur des énoncés contenant le prédicat de vérité.

Ketland et Shapiro considèrent que le raisonnement ci-dessus est parfaitement légitime et correspond d'ailleurs à la pratique courante des mathématiciens et logiciens. C'est pourquoi, selon eux, une théorie adéquate de la vérité doit pouvoir en rendre compte, c'est-à-dire qu'elle doit permettre de le formaliser. Voici les passages pertinents de articles de Ketland et Shapiro dans lesquels ils introduisent ce type d'arguments sémantiques :

Tout d'abord Ketland :

> « Retrouver le ou les passages pertinents »

Shapiro pour sa part se concentre principalement sur les arguments donnés dans le cadre des réflexions accompagnant certaines interprétations du théorème de Gödel. Néanmoins, on retrouve au coeur de ces raisonnements le même type de démonstration d'un principe de réflexion (*cf.* le passage mis en gras par nous dans l'extrait ci-dessous) :

> « Revenons à notre théorie de l'aritthmétique A et son énoncé de Gödel G (ou Con). Supposons qu'un professeur de logique affirme que G est vrai, et qu'une étudiante décontenancée demande une explication. L'étudiante croit le professeur sur parole quand il affirme que G est vrai, mais elle veut qu'on lui montre pourquoi cela est vrai. L'étudiante attend quelque chose comme une preuve convaincante ou ayant un pouvoir explicatif. La réponse naturelle est de faire remarquer que **tous les axiomes de A sont vrais et que les règles d'inférences préservent la vérité. Donc, tous les théorèmes de A sont vrais**[74]. Il s'ensuit que « $0 = 1$ » n'est pas un théorème de A et que par conséquent A est consistante. L'énoncé de Gödel est équivalent à la consistance de A. Il me semble que cette version informelle de Con et G est une *explication*[75] aussi bonne qu'une explication peut l'être. L'argument montre pourquoi G est vrai [...] Notre étudiante voulait savoir pourquoi G est vrai —ou pourquoi G est une conséquence— et le passage par la notion de vérité fournit l'explication. » Shapiro, 1998b, p. 505.

Puis, Ketland :

74. Nous soulignons.
75. Italiques de l'auteur.

> « Nous pouvons certainement “reconnaître” qu'un énoncé de Gödel G pour T est vrai (sous l'hypothèse que T elle-même est vraie), mais notre connaissance de sa vérité ne découle *pas* de dérivations formelles correctes *à l'intérieur* de la théorie T à laquelle il s'applique. [...] Dès lors, comment “reconnaissons-nous la vérité” de G ? [...] G est *déductible* de la théorie renforcée : à savoir, T *plus la théorie Tarskienne de la vérité standard pour le langage de* T.
>
> [...] Si j'ai raison, notre capacité à reconnaître la vérité des énoncés de Gödel implique une théorie de la vérité (celle de Tarski) qui *transcende de manière significative les théories déflationnistes.* » KETLAND, 1999, p. 87–88, italiques de l'auteur

De ces preuves sémantiques permettant de justifier les principes de réflexion, Shapiro et Ketland tirent une seconde contrainte portant sur les théories formelles de la vérité :

Contrainte de Réflexivité. [76] Soit T une théorie de base exprimée dans un langage $\mathcal{L}$ qui ne contient pas de vocabulaire sémantique. On suppose que T est suffisamment riche pour exprimer sa propre syntaxe. Supposons qu'on enrichisse $\mathcal{L}$ par un prédicat de vérité « Vr » et qu'on ajoute à T des axiomes pour la vérité, formulés dans $\mathcal{L}' \supseteq \mathcal{L} \cup \{Vr\}$ et notés $V(T)$, alors la théorie étendue $T \cup V(T)$ doit permettre doit permettre de prouver l'énoncé « tous les théorèmes de T sont vrais ».

Remarquons dès à présent que, sans même parler de conservativité et de son incompatibilité éventuelle avec cette contrainte de réflexivité avancée par Shapiro et Ketland, la contrainte de réflexivité, à elle seule, pose déjà problème pour le déflationnisme en matière de vérité. En effet, c'est un résultat bien connu qu'une théorie de la vérité limitée aux seules instances de la CONVENTION **T** est trop faible pour permettre de prouver des énoncés généraux (1., 2., et 3.) qui apparaissent dans la preuve sémantique examinée par Shapiro et Ketland. [77] Plus généralement, une théorie de la vérité limitée à l'ensembles des **T**-équivalences est bien souvent insuffisante pour prouver des énoncés universels dont elle permet pourtant d'établir les instances. Par exemple, si on considère une théorie de base T énoncée en logique classique du premier ordre et capable de coder sa propre syntaxe, alors l'extension $T \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \ / \varphi \in \mathcal{L}_T\}$ suffit à prouver chaque instance du

76. Cette terminologie est empruntée de Ketland. KETLAND, 2005 parle ainsi d'argument de la réflexion contre le déflationnisme.

77. Ce résultat remonte à TARSKI, 1935. Pour un exposé plus précis des résultats techniques, voire la section suivante.

tiers-exclu pour les énoncés du langage $\mathcal{L}_T$, mais elle est incapable de démontrer l'énoncé de la loi générale correspondante.[78] De ce point de vue, les énoncés universels (1., 2., et 3.) ne font pas exception : une théorie déflationniste formalisée par la collection infinie des **T**-équivalences suffit à prouver de chaque axiome de T qu'il est vrai[79], mais elle est incapable d'établir l'énoncé général correspondant : « tous les axiomes de T sont vrais ». Cette faiblesse relative des théories aléthiques formalisées au moyen de l'ensemble des **T**-équivalences conduisait déjà Tarski à les considérer comme insuffisantes et à les rejeter comme inadéquates[80] . Ce point concernant l'inanité des **T**-équivalences pour prouver des généralisations universelles a aussi déjà été soulevé contre les déflationnistes dès 1993 par Anil Gupta (*cf.* Gupta, 1993[81]).

Ainsi, le problème soulevé par la réflexivité peut être vu comme un cas particulier d'un phénomène plus large touchant les théories déflationnistes, à savoir leur incapacité à établir certains énoncés généraux. Toutefois, l'analyse de Shapiro et Ketland a cet avantage de mettre en lumière un argument précis tiré de la pratique scientifique et dans lequel la notion de vérité semble jouer un rôle crucial. À ce titre, une simple exigence de fidélité à l'usage courant pourrait suffire à montrer l'inadéquation d'une théorie limitée à la seule collection infinie des **T**-équivalences : si la contrainte de réflexivité doit bien être remplie par toute théorie satisfaisante de la vérité, alors les théories de la vérité sur lesquelles s'appuient les déflationnistes[82] apparaissent comme inadéquates, simplement parce qu'elles sont trop faibles pour formaliser les arguments sémantiques employés en mathématiques, ou, pour le dire autrement, simplement parce qu'elles sont trop faibles pour rendre compte d'un emploi ordinaire et légitime du prédicat de vérité. Il convient au contraire de leur préférer des théories formalisées plus fortes, qui elles permettront

78. De manière plus détaillée : dans $T' = T \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi / \varphi \in \mathcal{L}_T\}$, pour tout énoncé φ du langage de T, on peut prouver ce qui suit :

1. $T' \vdash \varphi \vee \neg\varphi$
2. $T' \vdash Vr(\ulcorner\varphi \vee \neg\varphi\urcorner) \leftrightarrow \varphi \vee \neg\varphi$
3. Donc, $T' \vdash Vr(\ulcorner\varphi \vee \neg\varphi\urcorner)$

En revanche, T' est trop faible pour prouver l'énoncé général de la loi du tiers-exclu : $T' \nvdash \forall x(\exists y Sen(y) \wedge x = conj(y, neg(y)) \rightarrow Vr(x))$.

79. Pour φ un axiome de T, trivialement $T \vdash \varphi$. D'où, tout aussi trivialement, $T' \vdash Vr(\ulcorner\varphi\urcorner)$. En revanche, $T' \nvdash \forall x(Ax_T(x) \rightarrow Vr(x))$.

80. Voyez la section sur Tarski 1.1.4 dans notre chapitre de présentation historique du déflationnisme.

81. Sans avoir rencontré à notre connaissance de réponse parfaitement satisfaisante de la part des déflationnistes.

82. Plus précisément, toutes les variantes déflationnistes (nombreuses voire majoritaires) qui s'appuient sur une théorie de la vérité limitée aux **T**-équivalences , comme par exemple la théorie minimale d'Horwich 1998, le décitationnisme de Field 1994 ; 1994 ou celui de Quine 1970.

de mener à bien l'argument sémantique et de justifier le principe de réflexion. De telles théories existent. La plus célèbre d'entre elles est sans doute la théorie tarskienne de la vérité qui s'appuie sur des clauses récursives compositionnelles.

Cependant, l'argument de Ketland et Shapiro ne s'arrête pas là. La contrainte de réflexivité prend une force toute particulière lorsqu'on la met en regard de la contrainte de conservativité et que l'on prend en compte les phénomènes d'incomplétude gödeliens.

3.2.3 Incomplétude, réflexivité et conservativité

Rappelons que le premier théorème d'incomplétude établit que toute théorie formelle T qui est cohérente et qui contient un minimum d'arithmétique est incomplète, au sens où il existe des énoncés du langage $\mathcal{L}_T$ de T qui ne sont ni démontrables ni réfutables dans T.[83] Il existe plusieurs manières de démontrer ce célèbre résultat. Toutefois, la plus répandue consiste à employer un raisonnement par diagonalisation qui permet de construire explicitement un énoncé indémontrable dans T.[84] Cet énoncé de Gödel pour T, généralement noté G_T, s'obtient en appliquant un théorème de point fixe à un prédicat de prouvabilité pour T, de sorte que :

$$T \vdash G_T \leftrightarrow \neg\exists x Dem_T(x, \ulcorner G_T \urcorner)$$ [85]

Ainsi, G_T est (prouvablement) équivalent à un énoncé affirmant : « G_T n'est pas prouvable dans T ». À partir de cette construction, on montre assez facilement que si T est cohérente, alors G_T n'est pas démontrable dans T. Si, de plus T est 1-cohérente, alors $\neg G_T$ n'est pas démontrable dans T.

Le second théorème d'incomplétude renforce le résultat précédent en montrant que si T est cohérente, contient un minimum d'arithmétique, et si d'autre part, le prédicat de prouvabilité $Dem_T(x, y)$ satisfait certaines contraintes naturelles,[86] alors T ne

83. Autrement dit, il existe au moins un énoncé φ du langage $\mathcal{L}_T$ tel $T \nvdash \varphi$ et $T \nvdash \neg\varphi$. En réalité, on peut même montrer qu'il existe une infinité de tels énoncés. Ces énoncés sont dit indécidables dans T.

84. Signalons que certaines démonstrations du premier théorème d'incomplétude ne passent pas par la construction explicite d'un énoncé indécidable (nous pensons ici, notamment, aux preuves qui s'appuient sur la théorie de la calculabilité et sur l'indécidabilité du problème de l'arrêt). Néanmoins, les démonstrations de ce type ne s'accompagnent pas des arguments « sémantiques » qui nous intéressent ici.

85. Pour plus de détails concernant le sens de ces formules et leur construction, voir ce qui suit ainsi que l'appendice technique.

86. Dans la plupart des démonstrations pleinement rédigées du second théorème d'incomplétude, on suppose que le prédicat de prouvabilité satisfait les propriétés suivantes, dites de Bernays-Löb : Pour tous énoncés φ, ψ de $\mathcal{L}_T$,

peut prouver sa propre cohérence. Plus précisément, le second théorème d'incomplétude montre qu'un certain énoncé du langage $\mathcal{L}_T$,

$$\neg\exists x Dem_T(x, \ulcorner 0 = 1 \urcorner),$$

que l'on note habituellement $Con(T)$ et qui exprime la cohérence de T, n'est pas démontrable dans T.

Ces résultats techniques étant rappelés, le point qui nous intéresse ici est le suivant : dès lors que l'on étend T par un principe de réflexion, les énoncés G_T et $Con(T)$ deviennent trivialement démontrables. Donc toute théorie assez forte pour prouver le principe de réflexion permettra de prouver ces énoncés. Ainsi, comme l'illustre l'extrait de Shapiro, 1998b qui suit le passage que nous avons souligné en gras dans la citation ci-dessus, les arguments sémantiques évoqués par Shapiro et Ketland se poursuivent pour donner une preuve de G_T et de $Con(T)$ [87]

1. Tous les axiomes de T sont vrais
2. Les règles d'inférence de T préservent la vérité
3. Donc, tous les théorèmes de T sont vrais.
4. Or, « $\neg(0 = 1)$ » est un théorème de T
5. Donc, « $\neg(0 = 1)$ » est vrai (par 4. et Vr-intro) [88]
6. Par conséquent, « $(0 = 1)$ » n'est pas vrai. (d'après 5. et les propriétés du prédicat de vérité) [89]

---

1. Si $T \vdash \varphi$, alors $T \vdash \exists x Dem_T(x, \ulcorner \varphi \urcorner)$
2. $T \vdash \exists x Dem_T(x, \ulcorner \varphi \to \psi \urcorner) \to (\exists x Dem_T(x, \ulcorner \varphi \urcorner) \to \exists x Dem_T(x, \ulcorner \psi \urcorner))$
3. $T \vdash \exists x Dem_T(x, \ulcorner \varphi \urcorner) \to \exists x Dem_T(x, \ulcorner \exists x Dem_T(x, \ulcorner \varphi \urcorner) \urcorner)$

87. On suppose donc dans l'argument qui suit :
    1. que T est cohérente et contient assez d'arithmétique pour coder sa propre syntaxe et vérifier les deux théorèmes d'incomplétude.
    2. que T est étendue par une théorie pour la vérité $V(T)$ qui est adéquate au sens de Tarski, c'est-à-dire qu'elle a pour conséquence toutes les **T**-équivalences .
    3. que $T \cup V(T)$ est assez forte pour prouver le principe de réflexion : $T \cup V(T) \vdash \forall x(\exists y Dem_T(y, x) \to Vr(x))$

La preuve qui suit s'effectue alors dans $T \cup V(T)$.

88. Nous disons par 4. et Vr-intro, mais nous aurions pu dire par 3. et 4. Cependant, ce n'est pas nécessaire. Remarquez que, pour tout $\mathcal{L}_T$-énoncé φ, dès lors que $T \vdash \varphi$, et que la théorie étendue au moyen d'axiomes pour la vérité prouve toutes les **T**-équivalences, on obtient aussitôt, $Vr(\ulcorner \varphi \urcorner)$. En l'espèce, $T \vdash \neg(0 = 1)$ et $T \cup V(T) \vdash Vr(\ulcorner \neg(0 = 1) \urcorner) \leftrightarrow \neg(0 = 1)$ implique que $T \cup V(T) \vdash Vr(\ulcorner \neg(0 = 1) \urcorner)$. Le schéma de réflexion 3. n'est donc pas indispensable pour dériver 5.

89. Là encore, le simple fait que l'extension aléthique renferme toutes les **T**-équivalences suffit pour

7. D'où, « $(0 = 1)$ » n'est pas un théorème de T, (par 6. et 3.)

   i.e. T est cohérente. [90]

À quoi nous pouvons ajouter, dès lors que l'équivalence de G_T et $Con(T)$ est généralement facilement prouvable dans T :

8. G_T. [91]

Les arguments sémantiques évoqués pas Shapiro et Ketland et qui s'appuient sur des principes de réflexion, sont couramment employés pas les mathématiciens et les logiciens, même s'ils se présentent souvent sous la forme de commentaires informels, [92] plutôt qu'au moyen d'une théorie de la vérité parfaitement articulée et formalisée. À tel point que l'on peut être tenté de parler à leur sujet d'interprétation majoritaire ou standard des théorèmes de Gödel. [93] Ces arguments semblent faire un usage crucial de la notion de

établir ce fait : si $Vr(\ulcorner 0 = 1\urcorner) \leftrightarrow 0 = 1$, alors $\neg Vr(\ulcorner 0 = 1\urcorner) \leftrightarrow \neg(0 = 1)$. Si en outre $Vr(\ulcorner \neg(0 = 1)\urcorner) \leftrightarrow \neg(0 = 1)$, alors $\neg Vr(\ulcorner 0 = 1\urcorner) \leftrightarrow Vr(\ulcorner \neg(0 = 1)\urcorner)$. Comme d'après (5.) $Vr(\ulcorner \neg(0 = 1)\urcorner$, il s'en suit que (6.) $\neg Vr(\ulcorner 0 = 1\urcorner)$.

90. Une fois formalisé, l'énoncé « « $(0 = 1)$ » n'est pas un théorème de T » n'est rien d'autre que $\neg\exists x Dem_T(x, \ulcorner 0 = 1\urcorner)$, c'est-à-dire $Con(T)$.

91. Remarquons au passage que, *si* dans la plupart des théories arithmétiques (comme par exemple l'arithmétique de Peano, ou même simplement l'arithmétique primitive récursive) l'équivalence $Con(T) \leftrightarrow G_T$ est facile à établir—ainsi $PA \vdash Con(PA) \leftrightarrow G_{PA}$—, dans les arguments sémantiques, il n'est pas *indispensable* de préalablement démontrer $Con(T)$ pour prouver G_T comme c'est le cas dans le raisonnement exposé ici. Une fois établi dans une extension sémantique assez forte le principe de réflexion $\forall x(\exists y Dem_T(y, x) \rightarrow Vr(x))$, on peut prouver directement (dans $T \cup V(T)$) l'énoncé G_T sans préalablement avoir dérivé $Con(T)$. Parvenu à 3. on poursuit au moyen d'un simple raisonnement par cas comme suit :

4' $T \vdash \neg\exists x Dem_T(x, \ulcorner G_T \urcorner) \rightarrow G_T$ (par construction de G_T)

5' $T \cup V(T) \vdash \exists x Dem_T(x, \ulcorner G_T \urcorner) \rightarrow Vr(\ulcorner G_T \urcorner)$ (instance du schéma de réflexion 3.)

6' $T \cup V(T) \vdash \exists x Dem_T(x, \ulcorner G_T \urcorner) \rightarrow G_T$ (par décitation)

7' $T \cup V(T) \vdash G_T$

C'est d'ailleurs, le plus souvent, ce type d'argument sémantique qu'on retrouve dans les commentaires accompagnant les preuves du théorème de Gödel, ne serait-ce que pour des raisons d'exposition, puisqu'en général on commence par exposer le premier théorème d'incomplétude et la construction de l'énoncé G_T ("vrai mais indémontrable dans T"), avant d'en venir à l'énoncé de la cohérence et au second théorème d'incomplétude. Si nous avons exposé dans le corps du texte l'argument 1-8 plutôt que celui donné dans la présente note, c'est avant tout pour rester fidèle à l'argument informel avancé par Shapiro (Shapiro, 1998b) dans l'extrait cité précédemment.

92. comme dans l'extrait de Shapiro, 1998b cité plus haut.

93. Ces arguments sont apparus dès la publication des résultats de Gödel, et on les retrouve très souvent de façon plus ou moins formalisée dans les explications accompagnant les théorèmes de Gödel. Voici quelques exemples de textes classiques où l'on pourra trouver une exposition de ce type d'arguments sémantiques : Nagel, Newman et Girard, 1989, Dummett, 1978 ou encore Tarski, 1969. Milne (2007, p. 219-223) contient également une liste (non exhaustive mais néanmoins impressionnante) de textes classiques et de manuels de logique qui s'inscrivent dans cette interprétation. Tennant, 2002, qui lui-même ne souscrit pas à ces arguments sémantique, parle ainsi de « l'interprétation orthodoxe » des théorèmes d'incomplétude. Lorsqu'il la critique et qu'il veut s'en démarquer, il la qualifie tout simplement

vérité et du principe de réflexion. Il peut donc sembler légitime d'exiger qu'une théorie correcte de la vérité ait quelque chose à dire à leur sujet. Comme nous l'avons rappelé, la collection des **T**-équivalences ne suffit pas pour formaliser ces arguments sémantiques. Mais ce qui est encore plus ennuyeux pour les déflationnistes, c'est que dans le cas d'une théorie de base T soumise aux phénomènes d'incomplétude, *toute* théorie de la vérité pour le langage de T qui satisfait la contrainte de réflexivité sera du même coup assez forte pour permettre de construire un argument sémantique prouvant $Con(T)$ ou G_T. Or, ces deux énoncés sont des énoncés exprimés dans le langage de T, et bien sûr, par les théorèmes d'incomplétude, ils ne sont pas démontrables dans T. Par conséquent, *toute* théorie de la vérité qui satisfait la contrainte de réflexivité devient aussitôt non conservative sur T.

Les résultats de logique mathématique montrent que la contrainte de conservativité et la contrainte de réflexivité sont tout simplement incompatibles. Il semble donc que les déflationnistes soient « coincés » : s'ils se cantonnent à une théorie « modeste » (*i.e.* conservative) de la vérité, comme par exemple la théorie limitée au seules **T**-équivalences , alors ils sont dans l'incapacité de formaliser les explications sémantiques justifiant les principes de réflexion et permettant d'établir la vérité des énoncés de Gödel. Mais en outre, il leur est impossible de renforcer leur théorie en y ajoutant d'autres axiomes de manière à pouvoir rendre compte des raisonnements sémantiques évoqués par Keltand et Shapiro, sous peine de briser la contrainte de conservativité, ce qui contredit la thèse selon laquelle le prédicat de vérité désigne une notion non « substantielle » et sans pouvoir explicatif. Tel est donc le dilemme insoluble auquel le déflationniste semble confronté.

L'attaque de Shapiro, 1998b et Ketland, 1999 contre le déflationnisme a fait couler beaucoup d'encre et suscité de nombreuses réactions depuis sa parution. Dans le chapitre suivant, nous examinons les diverses réponses et discussions qui ont suivi cet argument. Auparavant, dans la section qui suit nous rappelons les principaux résultats techniques sur lesquels le débat s'est appuyé. Cet exposé technique, ou ses rappels, nous semble indispensable car certains des arguments échangés dans la discussion reposent en effet sur une connaissance fine et détaillée de résultats logico-mathématiques qui, s'ils sont bien connus depuis Gödel et Tarski, n'ont pas fini de provoquer les commentaires philosophiques.

de « dogme substantialiste ».

3.3 Théories arithmétiques, théories sémantiques

Comme nous l'avons vu dans les deux sections précédentes, l'essentiel du débat s'articule autour de la faiblesse (entendez la conservativité) ou la force (entendez la capacité à prouver le principe de réflexion et à formaliser les arguments sémantiques développés dans le sillage des phénomènes d'incomplétude gödeliens) de telle ou telle théorie de la vérité qu'on souhaiterait ajouter à une théorie de base soumise aux théorèmes d'incomplétude. Nous considérerons donc une théorie formelle de base T exprimée dans un langage $\mathcal{L}_T$ qui ne contient pas de terminologie sémantique.[94] On supposera néanmoins que $\mathcal{L}_T$ contient assez de terminologie pour pouvoir parler de sa propre syntaxe et que T contient assez de ressources pour pouvoir servir de théorie syntaxique de son propre langage. Autrement dit T est assez forte pour prouver des choses comme : « x est (le code d') un énoncé du langage $\mathcal{L}_T$ », « x est (le code d') un énoncé qui est la conjonction de l'énoncé (codé par) y et de l'énoncé (codé par) z », « x est (le code d'un énoncé) dérivable à partir des énoncés (codés par) y et z au moyen de telle règle d'inférence », « x est (le code d') un théorème de T », *etc.* Pour cela, on peut supposer que T contient assez d'arithmétique pour pouvoir réaliser un codage[95] à la Gödel de la syntaxe de $\mathcal{L}_T$.[96] Quoi qu'il en soit de la manière dont on traite la syntaxe de $\mathcal{L}_T$, on supposera néanmoins que T contient assez d'arithmétique[97] pour être soumise aux phénomènes d'incomplétude gödeliens.

94. En particulier le vocabulaire de $\mathcal{L}_T$ ne contient pas de prédicat de vérité « Vr », pas plus que de prédicat de satisfaction « Sat » ou d'autres prédicats « sémantiques » à partir desquels on pourrait tenter de définir la vérité.

95. Une « arithmétisation », comme on dit aussi.

96. Soulignons cependant que ce passage par l'arithmétique n'est pas à strictement parler indispensable. On pourrait également imaginer que T et $\mathcal{L}_T$ ne sont pas un langage et une théorie censés « parler » des nombres, mais qu'ils permettent néanmoins (peut-être après avoir été convenablement enrichis) de formaliser directement la syntaxe, en ayant pour cela les ressources nécessaires (par exemple un symbole représentant chaque symbole primitif de $\mathcal{L}_T$, un opérateur de concaténation, de quoi définir les expressions bien formées du langage $\mathcal{L}_T$, de quoi représenter le fait que tel énoncé de $\mathcal{L}_T$ est dérivable dans T, *etc.* (Voir Tarski, 1935 pour un exposé détaillé de la manière dont ceci peut être réalisé)). Le fait, parfaitement standard chez les logiciens et au demeurant bien pratique, d'employer une théorie arithmétique comme théorie syntaxique en identifiant les objets syntaxiques à leurs codes numériques se justifie par le fait que les suites de symboles sur un alphabet fini ou dénombrable sont isomorphes aux nombres entiers (sur ce point voir Corcoran, Frank et Maloney, 1974).

97. Là encore, signalons que T peut ne pas être une théorie arithmétique à l'origine (voyez sur ce point notre rappel (section 3.1.3 page 226) de la construction « nominalistiquement acceptable » des théorèmes de Gödel réalisée par Shapiro, 1983 dans sa critique des thèses défendues par Field, 1980). Il faut simplement que T soit assez forte pour pouvoir représenter sa propre syntaxe et la propriété « être prouvable dans T ».

3.3.1 Une théorie arithmétique

3.3.1.1 Le point de départ

Pour fixer les idées et rester aussi simple que possible, nous suivrons la pratique habituelle des logiciens,[98] en partant d'une théorie T arithmétique.[99] Supposons donc que nous nous donnions un langage-objet du premier ordre $\mathcal{L}_T$ qui contient la terminologie nécessaire pour formuler une théorie de l'arithmétique mais qui ne contient pas de vocabulaire sémantique. Par exemple, on peut prendre pour $\mathcal{L}_T$ le langage de l'arithmétique de Peano. $\mathcal{L}_T$ contient un terme $\overline{n}$ pour chaque entier naturel n, par exemple obtenu par itération de la fonction successeur appliquée au symbole primitif désignant zéro $\overline{n} = \underbrace{SS\ldots S}_{n \text{ fois}}(\overline{0})$. Dans ce langage, nous formulons une théorie arithmétique de base T, disons par exemple une axiomatisation au premier ordre de l'arithmétique du genre de celle de Peano. On supposera que cette théorie de base T est récursive et qu'elle contient les ressources nécessaires pour représenter sa propre syntaxe, au moyen d'un codage « à la Gödel » qui associe à chaque suite de symboles de $\mathcal{L}_T$ un entier. Suivant l'usage nous noterons $\ulcorner\varphi\urcorner$ le numéro de Gödel de φ pour $\varphi \in \mathcal{L}_T$.[100] D'autre part, on supposera que ce codage se comporte « correctement » au sens où à chaque opération syntaxique sur les expressions de $\mathcal{L}_T$ correspond une opération récursive sur les codes. Par exemple, à l'opération syntaxique consistant à construire la conjonction de deux expressions de $\mathcal{L}_T$: $\varphi_1, \varphi_2 \mapsto (*\ \varphi_1 * \wedge * \ \varphi_2\ *)$ correspond une opération récursive sur les codes : $conj(\ulcorner\ \varphi_1\urcorner, \ulcorner\varphi_2\urcorner) = \ulcorner\varphi_1 \wedge \varphi_2\urcorner$, et de même pour les opérations syntaxiques plus complexes (comme la substitution, le remplacement *etc.*).

3.3.1.2 Représentabilité

Tout ceci assure que les propriétés « être (le code d') un axiome de T », « être (le code d') une formule obtenue en appliquant une règle d'inférence de T à telle et telle formules (codées par...) », « être (le code) d'une preuve dans T » sont des propriétés récursives sur les entiers. En outre, on supposera que T est une théorie arithmétique assez forte

98. Laquelle est d'ailleurs adoptée par tous les protagonistes du débat qui nous intéresse.

99. Ce choix est simplement commode pour l'exposition technique que nous entreprenons ici. Comme nous venons de le dire, n'importe quelle théorie contenant assez d'arithmétique (ou une interprétation de l'arithmétique) pour qu'apparaissent les phénomènes d'incomplétude ferait l'affaire.

100. Et pour faciliter la lecture, nous noterons ainsi également le terme (*i.e.* le numéral) de $\mathcal{L}_T$ qui désigne cet entier, bien qu'en toute rigueur nous devrions écrire : $\overline{\ulcorner\varphi\urcorner}$.

pour satisfaire certaines propriété de réprésentatabilité. [101] Pour ce qui nous concerne, disons que nous disposons d'un théorème de représentabilité dans T qui nous assure que toutes les fonctions et toutes les relations récursives sont réprésentables dans T au sens où si $R(n_1, \ldots, n_k)$ est une relation récursive sur les entiers, et si $f(n_1, \ldots, n_k) = m$ est une fonction récursive sur les entiers, alors il existe une formule $\phi_R(x_1, \ldots, x_k)$ et une formule $\phi_f(x_1, \ldots, x_k, y)$ de $\mathcal{L}_T$ telles que :

$$(i) \quad (n_1, \ldots, n_k) \in R \text{ ssi } T \vdash \phi_R(\overline{n_1}, \ldots, \overline{n_k})$$

$$(ii) \quad (n_1, \ldots, n_k) \notin R \text{ ssi } T \vdash \neg\phi_R(\overline{n_1}, \ldots, \overline{n_k})$$

$$(iii) \quad f(n_1, \ldots, n_k) = m \text{ ssi } T \vdash \forall y(\phi_f(\overline{n_1}, \ldots, \overline{n_k}, y) \leftrightarrow y = \overline{m})$$

Notons que des théories de l'arithmétique très faibles suffisent pour que ce théorème soit vérifié. Ainsi, l'arithmétique de Robinson $\mathcal{Q}$, qui ne contient pas de schéma d'induction, est néanmoins assez forte pour vérifier ce théorème et il en sera, bien entendu, de même de toutes les théories arithmétiques étendant $\mathcal{Q}$.

Sous ces hypothèses, on peut alors construire dans $\mathcal{L}_T$ les formules suivantes :

1. Une $\mathcal{L}_T$-formule, notée $Ax_T(\overline{m})$, qui exprime que m est le numéro de Gödel d'un énoncé qui est un axiome de T. C'est-à-dire que

   $T \vdash Ax_T(\overline{m})$ *si et seulement si* m est le code d'un axiome de T.

2. Une $\mathcal{L}_T$-formule, notée $Inf_T(\overline{p}, \overline{m}, \overline{n})$, qui exprime que p est le code d'une formule obtenue en appliquant une règle de dérivation aux formules codées respectivement par m et par n. C'est-à-dire que

   $T \vdash Inf_T(\overline{p}, \overline{m}, \overline{n})$ *ssi* p est le code d'une formule obtenue en appliquant une règle de dérivation aux formules codées respectivement par m et par n.

3. Une $\mathcal{L}_T$-formule, notée $Dem_T(x, y)$, qui représente un prédicat de prouvabilité pour T. C'est-à-dire que

   (a) $T \vdash Dem_T(\overline{m}, \overline{n})$ si et seulement si m est le numéro de Gödel d'une preuve dans T de la formule codée par n et

   (b) $T \vdash \neg Dem_T(\overline{m}, \overline{n})$ si et seulement si m n'est pas le code d'une preuve de la formule codée par n,

101. Ici, la terminologie varie grandement d'un auteur à l'autre (représentabilité forte, faible, bi-énumérabilité, *etc.*).

3.3.1.3 Incomplétudes

3.3.1.3.1 Premier théorème d'incomplétude : G1

La représentabilité (dans T) de la prouvabilité (dans T) est l'une des clés de voute de la démonstration du premier théorème d'incomplétude. La seconde est le théorème de point fixe, aussi appelé lemme de diagonalisation, que nous avons déjà évoqué dans la section précédente. Ce lemme nous dit que pour toute formule $\varphi(v)$ à une variable libre de $\mathcal{L}_T$, il existe une formule ϕ de $\mathcal{L}_T$ telle que $T \vdash \phi \leftrightarrow \varphi(\ulcorner \phi \urcorner)$. [102] En l'appliquant à la formule $\neg\exists x Dem_T(x, y)$, on obtient l'énoncé de Gödel pour T :

$$(Diag) : T \vdash G_T \leftrightarrow \neg\exists x Dem_T(x, \ulcorner G_T \urcorner)$$

Intuitivement, l'énoncé G_T est équivalent à un énoncé affirmant que l'énoncé « G_T » n'est pas prouvable dans T. De là, on montre facilement que *si* T est cohérente, alors G_T n'est pas prouvable dans T. En effet, supposons que

(H) 1. $T \vdash G_T$,

alors par $(Diag)$ il suit immédiatement que

2. $T \vdash \neg\exists x Dem_T(x, \ulcorner G_T \urcorner)$

Mais d'autre part, si $T \vdash G_T$, il existe un entier n qui code cette preuve. Autrement dit, on a

3. $T \vdash Dem_T(\overline{n}, \ulcorner G_T \urcorner)$, d'où
4. $T \vdash \exists x Dem_T(x, \ulcorner G_T \urcorner)$
5. $T \vdash \exists x Dem_T(x, \ulcorner G_T \urcorner) \wedge \neg\exists x Dem_T(x, \ulcorner G_T \urcorner)$ ce qui est absurde si T est cohérente.

D'où, déchargeant (H), $T \nvdash G_T$.

Pour montrer à présent que $\neg G_T$ n'est pas dérivable dans T, il faut renforcer quelque peu nos hypothèses sur notre théorie arithmétique T. On supposera qu'elle est ω-cohérente : une théorie est ω-cohérente si pour toute formule $\varphi(x) \in \mathcal{L}_T$ à une variable libre telle que $T \vdash \exists x \varphi(x)$, il existe au moins un entier naturel n, $T \nvdash \neg\varphi(\overline{n})$. Bien sûr, toute théorie qui est ω-cohérente est *a fortiori* cohérente, mais l'inverse n'est pas toujours vrai. Supposons donc que notre théorie T est ω-cohérente. Alors si

102. Là encore, soulignons que les conditions minimalement suffisantes portant sur une théorie arithmétique pour que celle-ci vérifie ce lemme sont assez faibles. Par exemple, toute théorie capable de représenter toutes les fonctions récursives satisfera ce lemme. C'est donc le cas de l'arithmétique de Robinson $\mathcal{Q}$ et toutes les théories arithmétiques au moins aussi fortes, comme par exemple PA.

(*H*) 1. $T \vdash \neg G_T$,

par (*Diag*) il suit immédiatement que

2. $T \vdash \exists x Dem_T(x, \ulcorner G_T \urcorner)$

Mais T est cohérente, et par conséquent d'après ce qui précède, il n'existe pas de preuve de G_T dans T. Autrement dit, par représentabilité

3. pour tout entier n, $T \vdash \neg Dem_T(\overline{n}, \ulcorner G_T \urcorner)$

(2.) et (3.) contredisent l'ω-cohérence de T. Au total, si T est ω-cohérente, $T \nvdash \neg G_T$.

3.3.1.3.2 Second théorème d'incomplétude : G2

Informellement, le second théorème d'incomplétude affirme que « toute théorie cohérente et contenant un minimum d'arithmétique ne peut pas prouver sa propre cohérence ». Plus formellement, si T est capable de représenter son propre système de preuve au sens que nous avons expliqué précédemment (*cf.* la section 3.3.1.2), alors la cohérence de T peut s'exprimer par l'énoncé $\neg\exists x Dem_T(x, \ulcorner 0 = 1 \urcorner)$ qui est un énoncé de $\mathcal{L}_T$, et qu'on note habituellement $Con(T)$. Il s'agit de montrer que T ne permet pas de dériver l'énoncé $Con(T)$. L'idée qui sous-tend la démonstration du second théorème d'incomplétude est celle qui avait déjà été déjà suggérée par Gödel dans son article de 1931, sans qu'il ait donné une preuve entièrement rédigée de ce second théorème. Le raisonnement qui sous-tend la preuve du premier théorème d'incomplétude s'appuie sur des notions élémentaires codables dans l'arithmétique. Sous certaines hypothèses, on peut donc le formaliser à l'intérieur de T elle-même, pour obtenir une « version formelle » du premier théorème : $T \vdash Con(T) \rightarrow G_T$.[103] Ceci fut réalisé pour la première fois par Hilbert et Bernays, 1939.

Pour alléger la notation, on notera $Thm(\ulcorner \varphi \urcorner)$ l'énoncé $\exists x Dem_T(x, \ulcorner \varphi \urcorner)$ et, quels que soient les énoncés φ, ψ de $\mathcal{L}_T$, on supposera que Thm vérifie les trois propriétés suivantes :

*D*1. Si $T \vdash \varphi$, alors $T \vdash Thm_T(\ulcorner \varphi \urcorner)$

*D*2. $T \vdash Thm_T(\ulcorner \varphi \rightarrow \psi \urcorner) \rightarrow (Thm_T(\ulcorner \varphi \urcorner) \rightarrow Thm_T(\ulcorner \psi \urcorner))$

*D*3. $T \vdash Thm_T(\ulcorner \varphi \urcorner) \rightarrow Thm_T(\ulcorner Thm_T(\ulcorner \varphi \urcorner) \urcorner)$

103. Ceci est une « version formelle » du premier théorème dans la mesure où, puisque G_T dit d'elle-même (*i.e.* est équivalent dans T à un énoncé exprimant) qu'elle n'est pas prouvable, $Con(T) \rightarrow G_T$ peut se voir comme une version formalisée de la conclusion du premier théorème d'incomplétude : si « T est cohérente alors G_T n'est pas prouvable ».

Tout d'abord, on montre le principe (A) suivant :
Pour toute formule $\varphi \in \mathcal{L}_T$,

$$(A)\ T \vdash Thm_T(\ulcorner \neg\varphi \urcorner) \leftrightarrow Thm_T(\ulcorner \varphi \to 0 = 1 \urcorner)$$

Démonstration.

1.	$T \vdash 0 \neq 1$	
2.	$T \vdash \neg\varphi \leftrightarrow (\varphi \to 0 = 1)$	
3.	$T \vdash Thm_T(\ulcorner \neg\varphi \leftrightarrow (\varphi \to 0 = 1) \urcorner)$	2., $D1$.
4.	$T \vdash Thm_T(\ulcorner \neg\varphi \urcorner) \to Thm_T(\ulcorner \varphi \to 0 = 1) \urcorner)$	3., $D2$.
5.	$T \vdash Thm_T(\ulcorner \varphi \to 0 = 1) \urcorner) \to Thm_T(\ulcorner \neg\varphi \urcorner)$	3., $D2$.
(A).	$T \vdash Thm_T(\ulcorner \neg\varphi \urcorner) \leftrightarrow Thm_T(\ulcorner \varphi \to 0 = 1) \urcorner)$	4. et 5.

□

Sous ces hypothèses, on peut alors raisonner comme suit :

1.	$T \vdash G_T \to \neg Thm_T(\ulcorner G_T \urcorner)$	par $(Diag)$
2.	$T \vdash Thm_T(\ulcorner G_T \to \neg Thm_T(\ulcorner G_T \urcorner) \urcorner)$	1. et D1.
3.	$T \vdash Thm_T(\ulcorner G_T \urcorner) \to Thm_T(\ulcorner \neg Thm_T(\ulcorner G_T \urcorner) \urcorner)$	2., D2. et modus ponens
4.	$T \vdash Thm_T(\ulcorner \neg Thm_T(\ulcorner G_T \urcorner) \urcorner) \to Thm_T(\ulcorner (Thm_T(\ulcorner G_T \urcorner) \to 0 = 1) \urcorner)$	par (A)
5.	$T \vdash Thm_T(\ulcorner G_T \urcorner) \to Thm_T(\ulcorner (Thm_T(\ulcorner G_T \urcorner) \to 0 = 1) \urcorner)$	3. et 4.
6.	$T \vdash Thm_T(\ulcorner G_T \urcorner) \to (Thm_T(\ulcorner (Thm_T(\ulcorner G_T \urcorner) \urcorner) \to Thm_T(\ulcorner 0 = 1 \urcorner))$	5. et D2.
7.	$T \vdash Thm_T(\ulcorner G_T \urcorner) \to Thm_T(\ulcorner (Thm_T(\ulcorner G_T \urcorner) \urcorner)$	D3.
8.	$T \vdash Thm_T(\ulcorner G_T \urcorner) \to Thm_T(\ulcorner 0 = 1 \urcorner)$	6. et 7.
9.	$T \vdash \neg Thm_T(\ulcorner 0 = 1 \urcorner) \to \neg Thm_T(\ulcorner G_T \urcorner)$	contraposition
$9'$.	$T \vdash Con(T) \to \neg Thm_T(\ulcorner G_T \urcorner)$	par déf. de $Con(T)$
10.	$T \vdash Con(T) \to G_T$	par 10. et $(Diag)$ à nouveau

À partir de là, on conclut facilement : si T prouvait $Con(T)$ alors, par (10), T prouverait G_T. Or, d'après le premier théorème d'incomplétude, on sait que c'est impossible.

Donc, $T \nvdash Con(T)$.

3.3.1.4 Remarques sur ces démonstrations d'incomplétude

À ce stade de notre parcours et avant d'en venir aux extensions aléthiques, il nous semble utile de formuler quelques remarques sur les démonstrations d'incomplétude que nous venons de rappeler. Ceci, pour attirer l'attention du lecteur sur un détail qui, bien qu'il n'ait pas eu d'impact direct sur les discussions de l'argument de la conservativité, n'est cependant pas sans importance.

Dans le débat sur le déflationnisme il est important de ne pas introduire subrepticement des notions sémantiques. Si l'on veut garder la trace du rôle éventuel que peut, ou que doit, jouer la notion de vérité, il faut donc bien séparer dans l'analyse de nos arguments ce qui appartient en propre à la vérité —et autres concepts sémantiques—, de ce qui relève des autres notions apparaissant dans nos raisonnements.

Les démonstrations d'incomplétude que nous avons retracées ci-dessus, en particulier celle du premier théorème d'incomplétude, s'appuient crucialement, outre les hypothèses de cohérence et d'ω-cohérence [104], sur la notion de réprésentabilité, c'est-à-dire sur la capacité de notre théorie de base T à prouver que certains entiers codant des énoncés vérifient certaines propriétés récursives. En particulier, la constructions de l'énoncé G_T et la preuve de son indécidabilité dans T, ne fait nullement appel à des notions sémantiques. À aucun moment, il n'est fait mention d'une quelconque hypothèse selon laquelle T serait vraie, ou selon laquelle tous les théorèmes, ou tels ou tels sous-ensembles des théorèmes de T seraient vrais. [105] La preuve du premier théorème d'incomplétude que nous avons donnée est ce qu'on peut appeler une version *syntaxique* de ce théorème.

Pour ce qui nous occupe, il faut bien la distinguer d'une autre version du premier théorème d'incomplétude qu'on peut appeler version *sémantique*. Cette autre preuve s'appuie aussi sur une construction par diagonalisation mais elle repose sur des hypothèses différentes de celle qui concerne la représentabilité dans T des propriétés récursives. Pour cette preuve *sémantique*, on suppose

104. Rappelons, pour mémoire, que l'hypothèse de l'ω-cohérence est inutilement forte. On peut la remplacer par ce qu'on appelle parfois la **1**-cohérence, qui est en fait une propriété d'ω-cohérence restreinte aux seuls énoncés Δ_0^0 : il n'existe pas de formule Δ_0^0, $\varphi(x)$ telle que $T \vdash \exists x \varphi(x)$ alors que pour tout n, $T \vdash \neg\varphi(\overline{n})$. Cela suffit puisque $Dem_T(x, \ulcorner G_T \urcorner)$ lui-même est Δ_0^0 (*i.e.* sans quantification non bornée (donc récursif)). C'est semble-t-il à Kreisel que revient le mérite d'avoir le premier noté cette possible amélioration consistant dans l'affaiblissement des hypothèses de départ sur T (Kreisel, 1957).

105. Quel que soit ce qu'on entend par là ...

1. que $\mathcal{L}_T$ a des ressources expressives suffisantes pour pouvoir « exprimer » ou « définir » [106] les propriétés récursives au sens suivant : pour toute propriété récursive sur les entiers, disons $P(n_1, \ldots, n_k)$, il existe une formule du langage $\mathcal{L}_T$, disons φ_P telle que

$$\varphi_P(\overline{n_1}, \ldots, \overline{n_k}) \text{ est vraie ssi les entiers } n_1, \ldots, n_k \text{ vérifient } P.$$

2. que T est récursive, au sens où ses axiomes et son système de preuves sont récursifs, de sorte qu'une fois fixé un codage gödelien de la syntaxe de $\mathcal{L}_T$, la relation « m est le code d'une preuve dans T de la formule codée par n » est *exprimable* dans $\mathcal{L}_T$. En d'autres termes, d'après 1. ci-dessus, il existe une formule de $\mathcal{L}_T$, notons la $Prouv(x, y)$, telle que $Prouv(\overline{m}, \overline{n})$ est vraie ssi m est le code d'une preuve dans T de la formule codée par n. Insistons sur le fait que nous nous appuyons ici sur la capacité expressive de $\mathcal{L}_T$ et non pas sur la puissance de réprésentabilité de T : T peut très bien être incapable de dériver $Prouv(\overline{m}, \overline{n})$, alors même que m est le code d'une preuve de la formule codée par n. [107]

3. que T est « fiable » [108] au sens tout ce qu'elle prouve est vrai. [109]

Sous ces hypothèses (1-3), on peut construire un énoncé G_T de $\mathcal{L}_T$ tel que G_T est vrai ssi $\neg\exists x Dem_T(x, \ulcorner G_T \urcorner)$ l'est. [110] Autrement dit G_T est équivalent à une formule qui énonce que G_T n'est pas prouvable dans T. Dès lors, si G_T (qui est vrai si et seulement si $\neg\exists x Dem_T(x, \ulcorner G_T \urcorner)$ est vrai, autrement dit si et seulement si G_T n'est pas prouvable dans T) était prouvable dans T, la théorie T prouverait un théorème faux. Mais par hypothèse (3.), T est fiable. Donc, G_T n'est pas prouvable dans T. Par conséquent, G_T est vrai. Donc, $\neg G_T$ est faux. Et donc, T ne prouve pas $\neg G_T$ non plus. [111]

Bien sûr, on pourrait être en droit de demander ce que les déflationnistes auraient à dire à propos des notions sémantiques apparaissant dans *cette démonstration là* du premier théorème d'incomplétude. La notion de vérité employée lorsqu'on suppose que

106. La terminologie varie selon les auteurs.
107. Comparez avec les hypothèses de 3.3.1.2.
108. ou « correcte », ou « T est une théorie vraie pour l'arithmétique », ...
109. Bien entendu, cette dernière hypothèse donnée ici de manière informelle, n'est rien d'autre que le schéma de réflexion. Dans une théorie formalisée contenant un prédicat de vérité pour $\mathcal{L}_T$, on pourrait l'exprimer par (Ref).
110. À nouveau, insistons sur le fait que la construction de G_T s'appuie sur les capacités expressives de $\mathcal{L}_T$ et se fait « en dehors de T ». Il n'est nullement supposé que $T \vdash G_T \leftrightarrow \neg\exists x Dem_T(x, \ulcorner G_T \urcorner)$.
111. Pour une exposition détaillée des diverses versions des diverses preuves des théorèmes de Gödel qui, notamment, insiste particulièrement sur cette distinction entre preuves *sémantiques versus syntaxiques*, voir Smith, 2013.

T est fiable au sens où tous ses théorèmes sont vrais, ou lorsqu'on part de l'hypothèse que $\mathcal{L}_T$ permet d'exprimer ou de définir les propriétés récursives dans la mesure où il existe une $\mathcal{L}_T$-formule qui sera vraie ssi la propriété est vérifiée, cette notion de vérité est-elle substantielle ? Joue-t-elle un rôle explicatif dans cette preuve *là* du premier théorème d'incomplétude ? Comment doit-on l'analyser ? Lorsqu'ici on dit « vrai », qu'entend-on au juste par là ? vrai *simpliciter* ? vrai dans $\mathbb{N}$? vrai au sens où : un énoncé $\ulcorner\varphi\urcorner$ est vrai ssi φ ? Si on devait donner une version entièrement formalisée de cette démonstration, une théorie formelle de la vérité acceptable selon les canons déflationnistes suffirait-elle à rendre compte du raisonnement poursuivi ? Peut-être, peut-être pas ; nous ne voulons pas préjuger cette question. [112]

Le problème c'est qu'ici les notions et les arguments sémantiques sont déjà présents dès la preuve de l'indécidabilité de G_T ; ils y sont en quelque sorte mélangés, ce qui, peut-être, rend difficile de distinguer ou de mesurer la force explicative qui peut leur être attribuée en propre. Par contraste, remarquez la preuve *syntaxique* précédente est sans conteste parfaitement acceptable pour un déflationniste, puisqu'aucune notion sémantique n'y est ne serait-ce que mentionnée et que seules interviennent des notions syntaxiques comme la prouvabilité et la réprésentabilité. Méthodologiquement, il est donc plus simple de prendre comme point de départ la preuve syntaxique, *i.e.* d'admettre que la démonstration de l'indécidabilité de G_T ne fait pas usage de notions sémantiques. Cette preuve ne nous dit rien quant à la vérité de G_T ou concernant la fiabilité de T [113] Ce n'est que dans un deuxième temps que nous étendrons notre théorie T et son langage $\mathcal{L}_T$ au moyen d'axiomes pour la vérité. Ceci permettra d'isoler ce qui revient en propre à la vérité, et d'évaluer précisément, du moins espérons le, sa « substantialité » ou l'étendue de son « pouvoir explicatif ». En somme, nous prenons acte du fait que les notions sémantiques, en particulier la notion de vérité, ne sont pas consubstantielles à la démonstration des théorèmes d'incomplétudes, ou pour le dire autrement, qu'elles ne lui sont pas indispensables. Il existe une preuve des théorèmes d'incomplétude parfaitement accessible aux déflationnistes. [114] Ce n'est que par la suite, lorsqu'il s'agit d'établir la

112. En prenant comme point de départ la version *syntaxique* des théorèmes de Gödel, nous pouvons en quelque sorte contourner cette question. Ou plutôt, nous reportons son examen à un deuxième temps, celui où nous introduisons les extensions aléthiques.

113. Par ailleurs, le fait que cette preuve ne suppose pas que tous les théorèmes de T sont vrais aura aussi son importance pour la discussion qui va suivre, voyez ....

114. Au demeurant, cette preuve *syntaxique* est également acceptable d'un point de vue finitiste à la Hilbert, ou du point de vue des mathématiques intuitionnistes ou constructives (l'argument diagonal employé dans la preuve syntaxique ne s'appuie en effet que sur des notions calculables et passe par une

vérité de G_T que la notion de vérité a, peut-être,[115] un rôle à jouer.[116]

Note historique : Sur un plan historique, il est intéressant de constater que les deux versions (sémantique et syntaxique) de la démonstration du premier théorème d'incomplétude sont déjà présentes dans l'article original de 1931. Dans la première section de son article, Gödel expose une version informelle de la preuve *sémantique* en s'appuyant sur une notion « intuitive », non formalisée de vérité[117] (voir Gödel, 1931, § 1, p. 147-151). Puis il consacre les sections suivantes (Gödel, 1931, § 2. et 3, p. 151-191) à établir le caractère (primitif) récursif d'une relation de prouvabilité, puis le caractère représentable de cette relation, avant d'en venir à la construction d'un énoncé de Gödel type G_T, prouvablement équivalent à un énoncé assertant sa non-prouvabilité. Autrement dit, il donne une version syntaxique de son premier théorème.[118] À notre connaissance Mostowski (1952) est la première publication à distinguer parfaitement les deux types de démonstration du premier théorème d'incomplétude et à proposer une version formalisée de la preuve *sémantique*. Il s'appuie pour cela sur une théorie tarskienne de la vérité.[119]

construction explicite de l'objet G_T). Cette acceptabilité de l'incomplétude était sans doute de la plus haute importance aux yeux de Gödel (sur ce point voir Feferman, 1998b).

115. Pour ce qui nous concerne c'est justement ce qui est en question.

116. Notez néanmoins, qu'un esprit moins conciliant pourrait protester et proposer le défi suivant aux déflationnistes : la preuve sémantique du théorème d'incomplétude est une bonne preuve, une preuve parfaitement acceptable du point de vue mathématique ! Une théorie adéquate de la vérité devrait donc pouvoir rendre compte du rôle que la notion de vérité semble tenir dans *cette* preuve.

Ceci étant, il semble à peu près clair que la discussion qui s'ensuivrait serait sans doute assez similaire à celle que nous examinerons dans le chapitre suivant et qui sépare les arguments sémantiques établissant G_T de la démonstration de son indécidabilité. Nous nous conformons donc au cadre dans lequel s'est développé la discussion sur le déflationnisme et les phénomènes gödeliens, en prenant pour point de départ la preuve *syntaxique* du premier théorème d'incomplétude.

117. Naturellement, à l'époque, c'est-à-dire avant Tarski, 1935, on ne disposait pas encore d'une formalisation claire du concept de vérité.

118. Sur les raisons qui ont poussé Gödel à passer sous silence, du moins pendant toute la période des années 1930, ses propres convictions philosophiques concernant la nature des mathématiques — notamment les notions transfinies ou la notion de vérité— et à mettre en avant la version syntaxique de son théorème plutôt que sa version sémantique, voir à nouveau Feferman, 1998b. Signalons cependant que la version *syntaxique* du premier théorème est indispensable pour pouvoir prouver le second théorème d'incomplétude à la manière de Hilbert et Bernays, 1939 (voyez l'usage important qui est fait de $(Diag)$ et donc de la représentabilité de $Prouv(x, y)$ dans cette démonstration p. 245).

119. Donc sur une théorie plus forte que la théorie décitationnelle formalisée par la collection des **T**-équivalences.

3.3.2 Extensions sémantiques

Nous sommes donc partis d'une théorie de base T, exprimée dans un langage $\mathcal{L}_T$, et contenant assez d'arithmétique pour pouvoir coder sa propre syntaxe, représenter la relation « m est le code d'une preuve dans T de l'énoncé codé par n », tant et si bien que T est soumise aux théorèmes d'incomplétude de Gödel (version syntaxique).

3.3.2.1 Un langage pour la vérité : $\mathcal{L}'$

À présent, supposons que l'on souhaite se donner une théorie de la vérité pour T et son langage. On commence donc par étendre $\mathcal{L}_T$ en un langage $\mathcal{L}'$ qui contient la terminologie requise pour formuler notre théorie de la vérité. $\mathcal{L}'$ contiendra au minimum un prédicat « Vr », que nous interpréterons comme un prédicat de vérité *pour* $\mathcal{L}$.[120] Dès lors que T possède déjà les ressources nécessaires pour formaliser sa propre syntaxe et que $\mathcal{L}'$ étend $\mathcal{L}_T$ (auquel cas, $\mathcal{L}'$ contiendra tous les énoncés de $\mathcal{L}_T$),[121] l'ajout d'un seul prédicat unaire « $Vr(x)$ », suffira pour pouvoir formuler une théorie déflationniste identifiée à l'ensemble infini des **T**-équivalences : $\{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi / \varphi \in \mathcal{L}\}$. Dans le cas d'une théorie T et d'un langage $\mathcal{L}_T$ pour l'arithmétique, on peut également tirer avantage du fait que le langage $\mathcal{L}_T$ contient un terme pour chaque objet dont parle la théorie[122] afin de formuler une théorie néo-tarskienne de la vérité pour $\mathcal{L}_T$ sous la forme d'une axiomatisation donnant les clauses récursives pour « Vr » sans passer par la satisfaction. Là encore, la seule ressource expressive supplémentaire que devra contenir $\mathcal{L}'$ pourra se limiter à l'ajout d'un nouveau prédicat « Vr » au vocabulaire de $\mathcal{L}_T$. Le débat suscité par l'argument de Ketland (1999) et Shapiro (1998a) s'est pour l'essentiel développé à partir d'une théorie de base T identifiée à l'arithmétique de Peano du premier ordre (PA) et augmentée de telle ou telle extension aléthique donnée sous la forme d'une axiomatisation d'un prédicat « Vr ». C'est donc essentiellement ce type d'extensions aléthiques, où $\mathcal{L}' = \mathcal{L} \cup \{Vr\}$, que nous allons examiner.

Ceci étant, il ne faut pas perdre de vue qu'il n'est pas exclu a priori que $\mathcal{L}'$ puisse aussi contenir éventuellement d'autres ressources lexicales nécessaires pour formuler notre

120. Rappelons que nous ne considérons ici que des théories formalisées pour un prédicat de vérité typé.
121. Si l'on préfère, on peut aussi dire, en termes plus ouvertement tarskiens, que le métalangage $\mathcal{L}'$ dans lequel on développe une théorie sémantique pour le langage objet $\mathcal{L}_T$ contient une traduction *homophonique* des énoncés de $\mathcal{L}_T$, c'est-à-dire une traduction telle que tout énoncé de $\mathcal{L}_T$ se voit traduit par lui-même dans $\mathcal{L}'$.
122. Précisément, un numéral $\overline{n} := \underbrace{SS\ldots S}_{n \text{ fois}}(\overline{0})$ pour chaque entier naturel n.

théorie aléthique. $\mathcal{L}'$ pourrait par exemple comprendre un prédicat « Sat » (pour Satisfaction) et le symbole d'appartenance « $\in$ » du vocabulaire de la théorie des ensembles. En toute généralité, sauf lorsque $\mathcal{L}_T$ contient un terme pour chaque éléments du « domaine » de T, ce détour par la satisfaction (d'une formule ouverte par une séquences infinies d'objets) sera même inévitable. [123]

Dans le vocabulaire étendu, $\mathcal{L}' \supseteq \mathcal{L} \cup \{Vr\}$, on peut à tout le moins **formuler** dans $\mathcal{L}'$ les généralisations suivantes concernant notre théorie de base T :

— L'énoncé

$$\forall x(Ax_T(x) \to Vr(x)),$$

qui exprime que « tous les axiomes de PA sont vrais ». On pourra le noter : $(AxVr)$

— L'énoncé

$$\forall x \forall y \forall z((Inf_T(x,y,z) \wedge Vr(y) \wedge Vr(z)) \to Vr(x)),$$

qui exprime que si x est (le code d'un) énoncé obtenu en appliquant une règle d'inférence à des énoncés vrais, alors il est vrai. Autrement dit, cet énoncé formalise l'assertion « les règles d'inférence préservent la vérité ». On pourra le noter : $(InfVr)$

— L'énoncé

$$\forall x(Thm(x) \to Vr(x)),$$

qui exprime que « tous les théorèmes de PA sont vrais », et qui n'est autre que le schéma de réflexion (Ref).

Les trois énoncés ci-dessus sont ainsi des traductions formelles dans le langage $\mathcal{L}'$ des prémisses qui apparaissent dans l'explication de la vérité de G_T donnée par Shapiro et Ketland (*cf.* page 233). Notez que les sous-formules $Ax_T(x)$, $Inf_T(x,y,z)$ et $Thm(x)$ sont des formules de $\mathcal{L}_T$, puisque par hypothèse nous avons supposé que T avait les

123. Rien de nouveau ici, voir Tarski, 1935. Précisons tout de même que nous ne voulons pas parler ici du problème de la définissabilité (explicite) dans $\mathcal{L}'$ de la notion de vérité pour $\mathcal{L}$, laquelle demanderait que (la métathéorie exprimée dans) $\mathcal{L}'$ renferme des ressources plus fortes que $\mathcal{L}$, disons un peu (plus) de théorie des ensembles (que T) ou des variables d'ordre supérieur à toutes celles de $\mathcal{L}$ (voir DeVidi et Solomon, 1999 ; Ray, 2005 ; Rouilhan, 1998). Quand bien même on se limiterait à une axiomatisation (sans définition explicite) dans $\mathcal{L}'$ d'une notion de vérité pour $\mathcal{L}$, bien souvent le passage par la satisfaction (d'une formule ouverte par une séquence infinie d'objets) est indispensable pour traiter le cas de la clause pour les quantificateurs —qu'on songe par exemple au cas d'une théorie exprimée dans un langage dénombrable mais censée porter sur un univers d'objets infini non-dénombrable.

ressources suffisantes pour représenter sa propre syntaxe. Le seul élément lexical nouveau apparaissant ici et n'appartenant pas à $\mathcal{L}_T$ est le prédicat de vérité « Vr ».

Dans $\mathcal{L}'$, on peut également donner des axiomes pour la vérité. Le passage par ces théories formalisées a au moins l'avantage de nous permettre de déterminer précisément quelles sont les théories de la vérité, exprimées dans $\mathcal{L}'$, qui nous permettent d'établir, c'est-à-dire de **prouver**, chacun des énoncés universels $(AxVr), (InfVr), (Ref)$ et qui donc peuvent rendre compte de manière formalisée du raisonnement décrit par Shapiro, ou à l'inverse quelles sont celles qui sont conservatives sur T. Nous allons donc nous livrer à présent à une comparaison de la force déductive de diverses théories de la vérité, couchées dans $\mathcal{L}'$.

3.3.2.2 Schémas d'axiomes

Avant d'exposer ces axiomatisations, rappelons, premièrement, que le cadre logique général dans lequel nous nous plaçons pour examiner les résultats qui suivent est la logique du premier ordre. Lorsque nous parlerons de conservativité nous pourrons donc entendre par là indifféremment une notion de conservativité syntaxique ou sémantique (puisqu'ici les deux notions coïncident). D'autre part, les débats autour de l'argument de la conservativité se sont quasi exclusivement développés en prenant comme théorie de base l'arithmétique de Peano et en mettant en regard les extensions aléthiques compositionnelles néo-tarskiennes et les extensions aléthiques formalisant la version décitationnelle du déflationnisme. Nous nous conformerons à cet usage en nous concentrant nous aussi sur ces deux types d'axiomatisation pour le prédicat « Vr » étendant un théorie de base T arithmétique. Enfin, un autre point important sur lequel il nous faut attirer l'attention concerne l'extension des schémas d'axiomes : si l'on considère une théorie de base T arithmétique formalisée au premier ordre, dans bien des cas cette théorie contiendra un schéma d'induction

$$SI : (\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(S(x))) \rightarrow \forall x\varphi(x)$$

identifié à l'ensemble de ses instances où φ est une formule de $\mathcal{L}_T$ —auquel cas T n'est pas finiment axiomatisée. Ainsi, chaque formule φ de $\mathcal{L}_T$ permet de construire un axiome d'induction dont la forme est donnée par le schéma ci-dessus. Mais, si on souhaite étendre T et que pour cela on enrichit le langage $\mathcal{L}_T$ en introduisant un nouveau vocabulaire, se pose alors la question de la manière dont on va traiter ce schéma d'axiomes. En effet,

supposons que $\mathcal{L}_T \subset \mathcal{L}'$ et que T' est une théorie exprimée dans le langage enrichi $\mathcal{L}'$ qui étend notre théorie de base T (autrement dit $T \subset T'$), lorsqu'on considère une formule φ' du vocabulaire étendu (*i.e.* $\varphi' \in \mathcal{L}'$) dans laquelle apparaissent des symboles qui n'étaient pas contenus dans $\mathcal{L}_T$, sommes-nous autorisés à prendre pour axiome l'énoncé $(\varphi'(0) \wedge \forall x(\varphi'(x) \rightarrow \varphi'(S(x))) \rightarrow \forall x\varphi'(x)$? Autrement dit, pouvons-nous appliquer les schéma d'axiomes de notre théorie d'origine aux formules du langage étendu $\mathcal{L}'$? Ce point est important car il peut avoir un impact sur les résultats de conservativité ou de non conservativité de T' sur T. Ce phénomène n'est pas propre aux schémas d'induction pour l'arithmétique. On retrouve exactement le même problème, et ses liens avec les résultats de conservativité, lorsqu'on s'intéresse à d'autres théories, généralement exprimées dans une logique du premier ordre, et contenant des schéma d'axiomes —comme par exemple les schémas d'axiome de compréhension en théorie des ensembles. Burgess et Rosen (1997) ont proposé la terminologie suivante qui est assez commode : lorsque les schéma d'axiomes est réservé aux seules formules du langage d'origine, on dit que le schéma est traité comme *liste*, quand au contraire on accepte d'appliquer le schéma aux formules du langage étendu, on dira que le schéma est traité comme *règle*.

3.3.2.3 Quatre extensions aléthiques

Voici donc les théories formalisées de la vérité qui ont été discutées par les protagonistes de la discussion de l'argument de la conservativité. Tout d'abord, on peut donner une axiomatisation pour « Vr » censée refléter les conceptions déflationnistes au sujet de la bonne axiomatisation du prédicat de vérité. On l'obtient en étendant simplement T au moyen de la collection infinie des **T**-équivalences .[124] Selon que l'on étend ou non le schéma d'induction, on obtient l'une ou l'autre version d'une théorie formelle déflationniste :

Définition. Extensions aléthiques décitationnelles :

(1) $T_d := T \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi : \varphi \in \mathcal{L}_T\}$ et le schéma d'induction est traité comme *liste*.

(2) $(T_d)^{+Ind} := T \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi : \varphi \in \mathcal{L}_T\}$ et le schéma d'induction est traité comme *règle*.

124. À quelques aménagements près, ceci est en droite ligne avec la théorie minimale défendue par Horwich, 1998b, ou avec les règles décitationnelles avancées par Field, 1994a,b. À ceci près que l'on remplace l'opérateur de nominalisation (sur les propositions) de Horwich, et l'opération de citation par mise entre guillemets de Field par un codage gödelien fournissant un nom pour chaque énoncé de $\mathcal{L}$.

Au delà des **T**-équivalences , on peut aussi souhaiter donner des axiomes supplémentaires gouvernant notre prédicat de vérité. En particulier, on peut donner des clauses récursives reflétant la manière dont la vérité des énoncés complexes dépend de la valeur de vérité des énoncés plus simples qui les composent. Ce type de clauses compositionnelles sont bien connues depuis Tarski, 1935. Là encore, selon la manière dont on étend ou non le schéma d'induction de T, on obtiendra l'une ou l'autre version d'une théorie formelle à la Tarski :

Définition. Extensions aléthiques tarskiennes :

(3) $T_{Tar} := T$ à laquelle on ajoute des clauses récursives « à la Tarski » et le schéma d'induction est traité comme *liste*.

(4) $(T_{Tar})^{+Ind} := T$ à laquelle on ajoute des clauses récursives « à la Tarski » et le schéma d'induction est traité comme *règle*.

Reste à expliquer ce qu'on entend précisément par « clauses récrusives à la Tarski ». Comme $\mathcal{L}_T$ contient un terme (à savoir $\overline{n}$) pour chaque objet de son domaine d'interprétation ($\mathbb{N}$), ces clause tarskiennes peuvent être énoncées sans qu'on soit contraint de faire un détour par une notion de satisfaction.

Voici une façon de le faire.[125] On part de $\mathcal{L}_T$. Une fois fixé un codage gödelien, à chaque opération syntaxique dans $\mathcal{L}_T$ consistant à construire un énoncé complexe au moyen de symboles logiques à partir d'énoncés plus simples, correspond une opération récursive sur les codes. Par exemple, il existe une fonction binaire $conj(x, y)$ récursive sur les entiers qui a tout couple de codes $\ulcorner\varphi\urcorner, \ulcorner\phi\urcorner$ de formules de $\mathcal{L}_T$ fait correspondre le code de la conjonction de ces deux formules $\ulcorner\varphi \wedge \phi\urcorner$:

$$\begin{array}{rrcl} conj: & \mathbb{N}\times\mathbb{N} & \longrightarrow & \mathbb{N} \\ & (x,y) & \mapsto & conj(x,y) \\ & (\ulcorner\varphi\urcorner,\ulcorner\phi\urcorner) & \mapsto & \ulcorner\varphi\wedge\phi\urcorner \end{array}$$

Étant donné les hypothèses que nous avons faites sur T, et comme $conj$ est une fonction récursive, il existe une formule de $\mathcal{L}_T$ qui permet de représenter cette fonction dans T. Quitte à passer par une extension définitionnelle de T, on peut donc supposer que $\mathcal{L}_T$ comporte un symbole de fonction binaire qu'on notera en plaçant un point au dessus du symbole de conjonction : « $\dot{\wedge}(x, y)$ » dont l'interprétation n'est autre que la fonction

125. Ceci est une construction classique et bien connue. Nous nous sommes inspirés ici de Halbach, 1999a

conj. On fait de même pour les autres fonctions récursives correspondant aux autres symboles logiques. [126] Ainsi, $\dot{\neg}(x)$ désigne la $\mathcal{L}_T$-fonction représentant dans T la fonction numérique qui à tout code d'une formule $\ulcorner\varphi\urcorner$ fait correspondre le code de la négation de cette formule $\ulcorner\neg\varphi\urcorner$. Pour le quantificateur universel, $\dot{\forall}(x, y)$ désignera la $\mathcal{L}_T$-fonction binaire qui étant donné en arguments $\ulcorner\varphi\urcorner$ le code d'une formule φ et i le code d'une variable v_i, donne en sortie le code de la formule $\forall v_i\varphi$. [127]

Pour faciliter la lecture, on oubliera les parenthèses et on placera les symboles fonctions binaires correspondants aux opérations syntaxiques impliquant les connecteurs binaires entre leurs deux arguments. De sorte que $\dot{\neg}(\ulcorner\varphi\urcorner)$ est noté $\dot{\neg}\ulcorner\varphi\urcorner$ et que $\dot{\vee}(\ulcorner\varphi\urcorner, \ulcorner\phi\urcorner)$ est noté $\ulcorner\varphi\urcorner\dot{\vee}\ulcorner\phi\urcorner$. De même, on écrira $\dot{\forall}v_i\ulcorner\varphi\urcorner$ au lieu de $\dot{\forall}(\ulcorner\varphi\urcorner, \ulcorner v_i\urcorner)$. Ces opérations sont représentables dans T : $T \vdash \dot{\neg}\ulcorner\varphi\urcorner = \ulcorner\neg\varphi\urcorner$ (... *et idem, mutatis mutandis,* pour les autre constantes logiques).

Par ailleurs, on note $En_{\mathcal{L}_T}$ la formule représentant l'ensemble des (numéros de Gödel des) formules closes de $\mathcal{L}_T$, $At_{\mathcal{L}_T}$ la formule représentant l'ensemble des formules atomiques de $\mathcal{L}_T$, et Vr_0 la formule représentant l'ensemble des formules atomiques vraies de $\mathcal{L}_T$. [128]

Enfin, l'opération qui, étant donné le code d'une formule φ, le code d'une variable v_i et un entier n donne en sortie le code de la formule $\varphi(\overline{n})$ obtenue en remplaçant dans φ toutes les occurrences libres de la variables v_i par le numéral $\overline{n}$ est elle aussi récursive. [129] On notera $x[\mathring{y}/v_i]$ la $\mathcal{L}_T$ expression désignant l'opération consistant à substituer

126. Autrement dit, $\dot{\wedge}(x, y)$ est une formule à deux variables libres qui représente la fonction $conj(x, y)$ dans T.

127. Autrement dit, pour tout code $n = \ulcorner\varphi\urcorner$ d'une $\mathcal{L}_T$-formule et pour tout $i = \ulcorner v_i\urcorner$ code d'une variable de $\mathcal{L}_T$,
$\dot{\forall}(n, i) = \dot{\forall}(\ulcorner\varphi\urcorner, \ulcorner v_i\urcorner) = \ulcorner\forall v_i\varphi\urcorner$.

128. Ici, on suppose donc que l'ensemble des (numéros de Gödel) des formules closes de $\mathcal{L}_T$, l'ensemble des (numéros de Gödel) des formules atomiques de $\mathcal{L}_T$ et l'ensembles des (numéros de Gödel) des formules atomiques vraies du langage $\mathcal{L}_T$ sont réprésentables dans T. Mais ceci n'est pas problématique : il est bien connu que ces ensembles de formules sont récursifs. Pour le montrer il faudrait détailler le codage de Gödel sur lequel nous nous appuyons, ce que nous ne ferons pas. Pour une démonstration en bonne et due forme de ces résultats, voir par exemple Boolos, Burgess et Jeffrey (2002, chaptire 23).

129. Plus précisément, la fonction ternaire suivante :

$$\begin{array}{rccc} \mathring{f}: & \mathbb{N}\times\mathbb{N}\times\mathbb{N} & \longrightarrow & \mathbb{N} \\ & (x, y, z) & \mapsto & \mathring{f}(x, y, z) \\ & (\ulcorner\varphi\urcorner, \ulcorner v_i\urcorner, n) & \mapsto & \ulcorner\varphi[v_i/\overline{n}]\urcorner \end{array}$$

où $\varphi[v_i/\overline{n}]$ est la formule obtenue en remplaçant par $\overline{n}$ toutes les occurrences libres de v_i dans φ

est récursive, donc représentable dans T. C'est ici qu'on tire avantage du fait que, premièrement, $\mathcal{L}_T$ contient un terme pour tous les entiers et, deuxièmement, du fait que l'opération qui associe à tout entier n le code du numéral $\overline{n}$ —*i.e.* : $n \mapsto \ulcorner\overline{n}\urcorner$— est récursive. Cette astuce permet de faire « entrer » la

le numéral $\overline{y}$ à la place de la variable v_i dans la formule x.

Tout ce qui précède est réalisé dans T et $\mathcal{L}_T$. À présent, on étend le langage $\mathcal{L}_T$ au moyen d'un prédicat « Vr » pour obtenir $\mathcal{L}' = \mathcal{L} \cup \{Vr\}$. Les clauses récursives à la Tarski s'énoncent alors comme suit dans $\mathcal{L}'$:

Définition. L'axiomatistion tarskienne Tar pour « Vr » est composée des axiomes suivants :

1. $\forall x \left(At_{\mathcal{L}_T}(x) \rightarrow (Vr(x) \leftrightarrow Vr_0(x))\right)$
2. $\forall x \left(En_{\mathcal{L}_T}(x) \rightarrow (Vr(\dot{\neg}x) \leftrightarrow \neg Vr(x))\right)$
3. $\forall x \forall y \left(En_{\mathcal{L}_T}(x \dot{\wedge} y) \rightarrow (Vr(x \dot{\wedge} y) \leftrightarrow (Vr(x) \wedge Vr(y)))\right)$
4. $\forall x \forall i \left(En_{\mathcal{L}_T}(\dot{\forall} v_i x) \rightarrow (Vr(\dot{\forall} v_i x) \leftrightarrow \forall y Vr(x[\mathring{y}/v_i])\right)$

Intuitivement, l'axiome 1. ci-dessus nous indique que si x est le code d'un énoncé atomique, alors x sera vrai si et seulement si x est un énoncé atomique vrai. L'axiome 2. garantit que si x est le code d'un énoncé de $\mathcal{L}_T$ la négation de x sera vraie si et seulement si x n'est pas vrai. Le troisième établit qu'une conjonction sera si et seulement si ses deux conjoints le sont. Et le dernier nous dit que tout énoncé de la forme $\forall v_i \varphi$ sera vrai si et seulement si tous les énoncés de la forme $\varphi(\overline{n})$, où $\overline{n}$ est un numéral désignant un entier naturel, le sont également.

Une fois ajoutée à T cette axiomatisation Tar pour la vérité, on obtient T_{Tar} si on considère les axiomes d'induction de T comme une liste (c'est-à-dire qu'on ne permet pas au nouveau vocabulaire propre à $\mathcal{L}'$ d'apparaître dans une induction) ou T_{Tar}^{+Ind} si l'on considère le schéma d'induction de PA comme un règle. Remarquez que contrairement à l'extension aléthique décitationnelle, cette axiomatisation au moyen des clauses récursives néo-tarskiennes donne une axiomatisation finie pour « Vr ».

Nous sommes maintenant en mesure d'exposer les résultats techniques pertinents pour notre discussion. [130]

Proposition 8. ***(*Ketland, *1999)*** *PA_d, et PA_d^{+Ind} sont toutes les deux conservatives sur PA. En particulier, elles ne permettent d'établir aucun des trois énoncés $(AxVr)$, $(InfVr)$, et (Ref).*

quantification à l'intérieur des codes et nous évite d'avoir à faire le détour par la notion de satisfaction dans notre axiomatisation pour le prédicat « Vr ».

130. Nous listons simplement ces résultats sans en donner les démonstrations mais en indiquant les références où on peut trouver ces dernières.

Ainsi, les théories décitationnelles ne peuvent même par établir (au sens de prouver) que tous les axiomes de T sont vrais. Notons cependant que, par la **T**-équivalence appropriée, elles suffisent bien à obtenir $Vr(\ulcorner\varphi\urcorner)$ pour tout axiome φ de T. Bien plus, si $T \vdash \varphi$ alors en appliquant la T-équivalence pour φ on obtient trivialement $Vr(\ulcorner\varphi\urcorner)$. Autrement dit, les extensions décitationnelles montrent bien *de* chaque axiome, ou *de* chaque théorème, qu'ils sont vrais. Mais elles ne suffisent pas à établir les généralisations correspondantes, à savoir que *tous les axiomes sont vrais* ou que *tous les théorèmes sont vrais*.

Les deux théories néo-tarskiennes sont plus fortes :

Proposition 9. *(***Halbach***, 2014)* *T_{Tar} et T_{Tar}^{+Ind} prouvent $(AxVr)$ et $(InfVr)$.*

Cependant T_{Tar} reste conservative sur T.

Proposition 10. *(***Halbach***, 1999a)* *T_{Tar} est conservative sur T et elle ne suffit pas pour prouver que tous les théorèmes de T sont vrais ($T_{Tar} \nvdash (Ref)$).*

Et enfin,

Proposition 11. *(***Halbach***, 2014)* *$T_{Tar}^{+Ind} \vdash (Ref)$.*

Corollaire 12. *T_{Tar}^{+Ind} n'est pas conservative sur T. En particulier elle permet de prouver G_T et $Con(T)$.*

Ainsi, une fois T étendue au moyen des clauses récursives à la Tarski (Tar), la théorie obtenue suffit pour établir que tous les axiomes de T sont vrais et que les règles d'inférence préservent la vérité. Néanmoins, seule T_{Tar}^{+Ind} permet de prouver le schéma de réflexion (Ref), c'est-à-dire de prouver que tous les théorèmes de T sont vrais. De là, il suit rapidement que T_{Tar}^{+Ind} montre que G_T et $Con(T)$ sont vrais, et donc qu'elle n'est pas conservative sur T. C'est ici que l'importance cruciale de la distinction entre comprendre un schéma d'axiomes comme *liste* ou comme *règle* apparaît. Une fois obtenu que tous les axiomes de T sont vrais, et que les règles d'inférence préservent la vérité, une simple récurrence sur la longueur des preuves montre que tous les théorèmes de T sont vrais. Cependant, pour pouvoir formaliser cette récurrence à l'intérieur de notre théorie étendue, on doit pouvoir recourir à une induction portant sur les formules contenant le prédicat « Vr ». Ceci n'est possible que si l'on a accepté d'étendre notre schéma d'induction au nouveau vocabulaire de $\mathcal{L}'$. Pour formaliser jusqu'au bout l'explication

de la vérité de l'énoncé G_T donnée par Shapiro et Ketland, T_{Tar} ne suffit pas, il nous faut T_{Tar}^{+Ind}. Pour Shapiro et Ketland ces résultats parlent sans doute en faveur d'une théorie tarskienne de la vérité avec induction étendue, la seule qui satisfasse la contrainte de réflexivité, plutôt que pour une axiomatisation limitée aux seules **T**-équivalences . Mais d'un autre côté, une telle théorie de la vérité n'est, selon eux, pas acceptable ou accessible pour un déflationniste puisqu'elle brise la contrainte de conservativité. Et comme ces deux contraintes sont logiquement incompatibles, toute tentative de la part du déflationniste pour les réconcilier semble sans espoir.

Après ce rappel des éléments techniques indispensables pour saisir les discussions de l'argument de la conservativité, nous sommes en mesure d'examiner plus précisément cet argument ainsi que les diverses réponses qu'il a suscitées. C'est ce que nous ferons au chapitre suivant. Quelles leçons philosophiques pouvons-nous tirer des faits logiques ci-dessus ? Plusieurs problèmes s'ouvrent devant nous. Tout d'abord, on peut se demander quelle est la « bonne » théorie de la vérité ? Toutes les théories que nous avons passées en revue satisfont CONVENTION **T**. Quels critères discriminants supplémentaires peut-on invoquer pour exercer une sélection parmi ces modèles concurrents ? Faut-il privilégier la modestie, sous la forme d'une contrainte de conservativité, ou au contraire la force explicative, sous la forme d'un critère d'adéquation à nos usages sémantiques « standard » tel que proposé par Shapiro et Ketland ? Puisque ces deux contraintes semblent incompatibles, le déflationniste peut choisir d'abandonner soit l'une soit l'autre de ces conditions. Mais il doit bien entendu justifier sa position. Une troisième possibilité a également été envisagée : le déflationniste peut tenter de dépasser l'antinomie face à laquelle Shapiro et Ketland semblent le placer en acceptant la contrainte de conservativité tout en reconnaissant qu'une « bonne » théorie de la vérité doit pouvoir rendre compte des arguments sémantiques à l'oeuvre dans l'établissement des schémas de réflexion et les explications de la vérité de G_T ou $Con(T)$, mais en montrant que la non-conservativité qui suit d'une telle théorie réflexive n'est pas réellement problématique : elle ne serait qu'un phénomène parasite, dont la vérité ne serait pas réellement responsable, et que par conséquent on ne pourrait pas interpréter comme la preuve ou l'indice du caractère « substantiel », ou du rôle explicatif, de la notion de vérité. En somme, le déflationniste peut soit refuser la première contrainte, soit refuser la seconde, soit montrer que leur incompatibilité n'est qu'apparente. Ces trois axes ont structuré les discussions qui ont suivi les publications de KETLAND (1999) et SHAPIRO (1998b). Dans ce qui suit, nous

les examinons tour à tour.

Chapitre 4

Discussion

Après ces longs prolégomènes qui nous ont permis d'expliquer en détails les hypothèses de l'argument de (Ketland, 1999 ; Shapiro, 1998b), ainsi que les outils techniques sur lesquels ils s'appuient, nous en venons enfin à l'examen des discussions que cet argument a suscitées.

Pour pouvoir suivre chaque étape de ces débats, voici un tableau récapitulatif de la force des extensions aléthiques qui seront examinées :

Extensions :	T_d	$(T_d)^{+Ind}$	T_{Tar}	$(T_{Tar})^{+Ind}$
Conservative sur T ?	oui	oui	oui	non
Prouve $(AxVr)$?	non	non	oui	oui
Prouve $(InfVr)$?	non	non	oui	oui
Prouve (Ref) ?	non	non	non	oui

Table 4.1 – Forces des extensions aléthiques

Nous rappelons également le squelette de l'argument :

(1) Une théorie déflationniste de la vérité doit être conservative.

(2) Une théorie adéquate de la vérité doit être réflexive.

(3) Toute théorie de la vérité réflexive sera non conservative.
Donc,

(C) Les théories déflationnistes de la vérité sont inadéquates

Comme nous venons de le rappeler à la fin du précédent chapitre, pour s'extraire du dilemme posé par Shapiro et Ketland, les défenseurs du déflationnisme peuvent adopter diverses stratégies. Ils peuvent tenter de réfuter l'une ou l'autre des deux premières prémisses de l'argument, c'est-à-dire contester l'une ou l'autre des deux contraintes censées peser sur une théorie de la vérité déflationniste. Ou bien encore ils peuvent tenter de montrer que l'incompatibilité logique de ces deux contraintes n'est pas si problématique qu'il y paraît. Les réponses mixtes combinant l'une ou l'autre des ces diverses voies de réponse ne sont évidemment pas exclues. Il est néanmoins commode de les examiner l'une après l'autre pour ordonner la discussion. Le plan de ce chapitre sera donc le suivant : dans la section qui suit nous revenons sur la contrainte de conservativité et tentons de montrer qu'elle s'impose bel et bien au déflationnisme. Dans un second temps, nous abordons directement le troisième type de réponse qui, pour l'essentiel, consiste à montrer que la non-conservativité d'une extension aléthique réflexive n'est pas imputable à la vérité et qu'elle ne remet pas en cause le caractère « non substantiel » ou « non explicatif » du prédicat de vérité. Là encore, nous argumentons contre ce type de réponse. Ce n'est qu'à la fin du chapitre que nous discutons la stratégie consistant à refuser la contrainte de réflexivité. La difficulté est alors de reconstruire les arguments sémantiques établissant G_T ou $Con(T)$ sans faire appel à une notion non-conservative de vérité, peut-être en proposant une autre justification des schémas de réflexion. Pour le dire d'emblée, si ce dernier type de réponse se heurte à divers problèmes que nous examinerons et si nous n'avons pas été convaincus, nous pensons néanmoins que c'est la stratégie la plus prometteuse pour le déflationniste, ou du moins celle qui débouche sur les questions les plus intéressantes.

4.1 Retour sur la conservativité

Dans l'abondante littérature parue en réponse à l'argument de Shapiro et Ketland, plusieurs voies, parfois contradictoires, ont été explorées. Néanmoins il est clair que l'examen motivé de la notion de conservativité appliqué à la vérité est évidemment central : dans quelle mesure est-elle une contrainte impérieuse pour le déflationniste ? Le moins que l'on puisse dire, c'est qu'il n'y a pas de consensus à ce sujet. Certains auteurs déflationnistes acceptent cette contrainte comme inévitable et essentielle à tout prédicat de vérité déflationniste. D'autres la récusent totalement. Nous commençons par passer en revue les postions des divers protagonistes sur cette question.

4.1.1 Conservativité : qu'en disent les déflationnistes ?

Dans quelle mesure la conservativité constitue une contrainte inévitable pour le déflationniste en matière de vérité ? Comme nous l'avons déjà rapidement évoqué, la plupart des auteurs, notamment ceux qui se réclament du déflationnisme, ne sont pas du tout d'accord entre eux sur cette question. Voici un rapide tour d'horizon des positions des divers protagonistes sur ce problème.

Parmi les défenseurs de la contrainte de conservativité appliquée au déflationnisme, on peut citer, outre Ketland et Shapiro que nous avons déjà évoqués, Leon Horsten, qui, semble-t-il, a été le premier à placer les déflationnistes devant une telle obligation. Dans un article paru en 1995, lorsqu'il examine les théories déflationnistes [1] Horsten écrit :

> La théorie minimale a pour conséquence qu'un prédicat de vérité devrait être conservatif sur une théorie donnée formulée sans prédicat de vérité (ou toutes autres notions sémantiques). (Horsten, 1995, p. 183)

Il ajoute ensuite dans une note de bas de page :

> Peut-être ai-je tort ici. Mais si tel est le cas, alors je ne vois pas en quoi consiste la neutralité de la notion de vérité selon le déflationniste. En tout cas, je considère que c'est un problème auquel les défenseurs des théories déflationnistes de la vérité devraient se confronter. (Horsten, 1995, note 15)

Ce type d'argument a été très souvent repris dans la littérature portant sur le déflationnisme aléthique. Par exemple, dans un article [2] où ils tentent de montrer que les fonctions expressives et explicatives de la vérité sont difficilement compatibles, si on les comprend à la manière déflationniste, Hyttinen et Sandu déclarent :

> Il est essentiel pour le déflationniste que l'extension [aléthique] [3] soit conservative sur PA, faute de quoi elle aurait des conséquences non tautologiques portant sur des questions qui ne concernent pas la vérité. Si tel était le cas, alors le déflationniste serait contraint d'admettre que le prédicat de vérité a

1. Plus précisément, la théorie minimale de Paul Horwich telle qu'elle est développée dans la première édition de Horwich (1990).
2. Hyttinen et Sandu (2004).
3. Hyttinen et Sandu font référence ici à l'extension aléthique de PA que nous avons notée T_d^{+Ind}, obtenue au moyen des seules **T**-équivalences (typées) : $PA \cup \{T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi : \varphi \in \mathcal{L}_{PA}\}$.

> ajouté un certain contenu à la théorie de base PA et qu'il est donc ontologiquement plus engageant qu'il n'était supposé l'être. (Hyttinen et Sandu, 2004, p. 414)

L'adoption d'une contrainte de conservativité n'est cependant pas forcément réservée aux critiques ou aux détracteurs du déflationnisme. Dans un article paru en 2002, [4] Neil Tennant, par ailleurs partisan d'une théorie antiréaliste de la vérité, [5] se fait l'avocat des déflationnistes et se propose de répondre en leur nom à l'argument de Ketland et Shapiro. Tennant semble [6] considérer que la conservativité est bel est bien essentielle à l'analyse déflationniste de la vérité. Selon lui, le déflationniste doit s'en tenir à T_d, la théorie purement décitationnelle, sans même élargir le schéma d'induction. Pour Tennant, en effet, le déflationniste considère que la vérité n'est pas une authentique propriété et par conséquent il ne doit pas accepter de l'introduire dans un schéma d'induction, outil servant justement à caractériser l'extension des (authentiques) propriétés. Tennant développe

4. Tennant (2002).

5.

> Je ne suis pas déflationniste. Je crois que la vérité et la fausseté sont substantielles. La vérité d'une proposition consiste en la possession d'une preuve constructive, ou vérifacteur [*truthmaker*] Tennant (2005, p. 89).

6. Un point de clarification bibliographique : dans un article de synthèse sur le débat autour de l'argument de la conservativité (Shapiro, 2003), Shapiro attribue cette position à Tennant :

> [Tennant] affirme que le déflationniste ne devrait pas accepter d'étendre le schéma d'induction aux formules qui contiennent le prédicat de vérité [. . .], parce que, pour le déflationniste la vérité n'est pas une authentique propriété. [. . .] Selon Tennant, la conservativité fait partie de l'essence de la position philosophique du déflationniste. Shapiro (2003, p. 111)

Shapiro cite comme source l'article de Tennant (2002), lequel est classé comme « à paraître » dans la bibliographie.

Néanmoins, dans l'article en question tel qu'il est paru dans *Mind*, on ne trouve nulle part de déclaration aussi explicite en faveur de la conservativité, pas plus qu'on n'y trouve de commentaire sur l'introduction de la vérité déflationniste dans l'induction. L'explication de tout ceci est peut-être la suivante : l'article de Shapiro est le texte d'une communication à un colloque qui s'est tenu en 1999, c'est-à-dire avant la publication de l'article de Tennant. Il est très vraisemblable que Shapiro a eu accès à une version antérieure de l'article, avant publication. Cette version contenait peut-être des développements supplémentaires concernant la contrainte de conservativité et l'induction sur la vérité, qui furent ensuite retirés et n'apparaissent donc plus dans la version publiée.

Quoi qu'il en soit, la teneur générale de l'article de Tennant, qui vise essentiellement à réfuter *la contrainte de réflexivité* au nom des déflationnistes, s'accorde bien avec la position que Shapiro lui attribue —peut-être un peu rapidement. Notons par ailleurs que dans un autre article plus récent (Tennant, 2010, p. 447), Tennant se classe parmi les auteurs souscrivant à une contrainte de conservativité pour la vérité, quoique non-déflationniste (Tennant se positionne comme un partisan d'un anti-déflationnisme conservatif, à l'intersection de la thèse selon laquelle la vérité est une propriété substantielle et de la thèse selon laquelle la théorie de la vérité doit être conservative sur le discours non sémantique (*cf.* la classification qu'il donne sous forme de tableau à la page 447 de son article)).

ensuite une argumentation visant à contourner la contrainte d'adéquation supplémentaire avancée par Ketland et Shapiro, en montrant que la vérité n'est pas indispensable pour établir G ou la cohérence de PA.

Si l'on se tourne ensuite vers les auteurs qui se déclarent ouvertement déflationnistes, là encore, les opinions varient. La position d'Hartry Field, l'un des principaux auteurs déflationnistes contemporains est subtile et difficile à saisir. Dans sa réponse à l'article de Shapiro, il ne se prononce jamais directement de manière explicite pour ou contre cette contrainte. Dans certains passages, il semble tout près de l'accepter, puisqu'il déclare :

> Étant donné que la vérité peut être ajoutée d'une manière qui engendre une extension conservative (même en logique du premier ordre), il n'est pas nécessaire d'être en désaccord avec Shapiro lorsqu'il dit : « la conservativité est essentielle au déflationnisme » (Field, 1999, p. 536, 1er §) [7]

D'un autre côté, Field insiste sur le fait que, même aux yeux d'un déflationniste, l'ajout de la notion de vérité augmente significativement le pouvoir expressif de notre langage et qu'

> il nous permet de faire des généralisations fécondes que nous ne pourrions pas faire autrement ; où par généralisation *féconde* je veux dire une [généralisation] qui a un impact sur des affirmations qui *n'impliquent pas la notion de vérité*. [8] Field (p. 533 1999, italiques de l'auteur)

Ce dernier passage déclare clairement que l'emploi de la notion de vérité peut « avoir un impact » en dehors du discours proprement sémantique ; ce qui suggère qu'une extension aléthique peut être non conservative, même au regard des canons déflationnistes.

Pour autant, le coeur de la réponse de Field s'appuie sur un résultat de conservativité pour ce qu'il appelle les « axiomes essentiels de la vérité ». Dans son article, Field se déclare en effet favorable à une version compositionnelle de la théorie de la vérité proche de T_{Tar}^{+Ind}. Pour répondre au dilemme posé par Ketland et Shapiro, son argumentation consiste à distinguer au sein de cette théorie, ce qu'il considère comme les

7. Au point qu'il a parfois été interprété comme acceptant (presque ?) la contrainte de conservativité (voir par exemple ce qu'en dit Halbach, qui considère Field comme étant « à deux doigts » — *come close to*— de se rallier à la contrainte de conservativité (Halbach, 2001b, p. 168)).

8. Reprenant la terminologie de Shapiro, Field remarque que pour un déflationniste si la vérité est, peut-être, *métaphysiquement maigre* elle n'est certainement pas *expressivement maigre* et que le prédicat de vérité peut servir à contracter des engagements concernant des sujets qui ne relèvent pas de la vérité, au-delà des engagements que nous pouvions prendre sans ce prédicat. (Field, 1999, p. 534)

axiomes relevant de la seule nature de la vérité, à savoir $\{Tar\}$, qui eux sont conservatifs [9] et les axiomes qui relèvent de notre compréhension des nombres entiers, les axiomes d'induction étendu à $\mathcal{L}'$, qui seuls permettent le passage de T_{Tar} à T_{Tar}^{+Ind} et provoquent la non conservativité. En filigrane, sans doute faut-il alors comprendre que la conservativité des « axiomes essentiels de la vérité » est importante pour le déflationnisme. En résumé, reconnaissant l'importance des arguments sémantiques évoqués par Shapiro [10], Field propose d'adopter une théorie axiomatique de la vérité *à la Tarski*, mais souligne que ces axiomes aléthiques sont conservatifs sur l'arithmétique de Peano et que la non-conservativité n'apparaît que lorsqu'on étend le schéma d'induction aux formules contenant le prédicat de vérité. Le déflationniste peut donc rester attaché à la conservativité (sur l'arithmétique de Peano) des axiomes purement sémantiques, tout en considérant que la non-conservativité de la théorie avec induction étendue est, malgré tout, compatible avec le caractère purement expressif de la vérité, telle que la conçoivent les déflationnistes.

Cette ligne d'argumentation a plus récemment été reprise et développée par Henri Galinon. [11] Le passage suivant nous semble illustrer de manière particulièrement claire la stratégie déployée :

> Je voudrais [...] soutenir que l'argument de Shapiro et Ketland, formulé en termes de non-conservativité, n'a pas la force que ces auteurs lui prêtent, en montrant que les phénomènes de non-conservativité sur lesquels ils attirent notre attention sont intuitivement compatibles avec l'hypothèse que le prédicat de vérité est simplement un outil expressif. Pour y parvenir, je me propose de prêter attention au statut épistémologique et logique des *schémas d'axiomes* dans ces théories. Une fois que les ambiguïtés de ce statut auront été levées, il deviendra plausible que la non-conservativité des théories de la vérité sur des théories comme l'arithmétique de Peano est un phénomène inoffensif pour la conception déflationniste. (Galinon, 2010, p. 155, italiques de l'auteur)

Là encore un accent particulier est mis sur le fait que la non-conservativité d'une extension aléthique tarskienne sur une théorie arithmétique n'apparaît que lorsque le schéma d'induction est traité comme règle, *i.e.* étendu au nouveau langage contenant

9. Comme nous l'avons rappelé dans le chapitre précédent, $T_{Tar} = T + Tar$ est conservative sur T.
10. Et donc, acceptant, semble-t-il, quelque chose comme la contrainte de réflexivité.
11. *cf.* Galinon (2010, en particulier chapitres 4 & 7).

le prédicat de vérité. Et ce phénomène est mis en avant à l'appui de l'idée que non-conservativité et caractère purement expressif sont compatibles.

L'autre figure majeure du déflationnisme contemporain, Paul Horwich, ne s'est pour sa part guère prononcé sur cette question de la conservativité. Dans *Truth* (Horwich, 1998b), il se déclare bien partisan d'une théorie *minimale* non compositionnelle, que, dans un version typée, on peut rapprocher de $T_d = \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi / \varphi \in \mathcal{L}\}$. Et certains passages de *Truth* semblent se prêter à une interprétation en termes de conservativité. Par exemple :

> Contrairement à la plupart des autres prédicats, « est vrai » ne sert pas à attribuer à certaines entités (*i.e.* assertions, croyances, *etc.*) un genre ordinaire de propriété —une caractéristique dont la nature sous-jacente rendra compte de ses relations avec d'autres ingrédients de la réalité. C'est pourquoi, à la différence des autres prédicats, on ne doit pas attendre de 'est vrai' qu'il prenne part à quelque théorie profonde concernant ce à quoi il réfère— une théorie qui articulerait des conditions générales de son application. (Horwich, 1998b, p. 2)

Si la vérité n'a pas de « nature sous-jacente » et que son attribution à telle ou telle entité ne doit pas permettre de rendre compte de ses relations avec les *autres* ingrédients de la réalité, on pourrait s'attendre à ce que l'ajout d' (théorie formalisant l'emploi d') un prédicat de vérité soit sans conséquences (nouvelles) sur le discours non sémantique ; autrement dit on pourrait s'attendre à ce qu'il soit conservatif.

Pour autant, dans une communication personnelle adressée à N. Tennant, Horwich déclare ne pas se sentir tenu par une contrainte de conservativité :

> ... cela ne m'inquiète pas le moins du monde que la théorie minimale ne soit pas conservative (à supposer qu'elle ne le soit pas) ; ... peut-être qu'une théorie conservative de la vérité serait, de ce point de vue, « plus déflationniste » que la mienne (comme le serait, peut-être, la thèse —que je rejette—selon laquelle il n'y a pas de propriété de vérité). Mais, et alors ? Pourquoi la théorie correcte de la vérité ne devrait-elle pas être non-conservative ? [...]
>
> [j'] attire l'attention sur le fait que (a) mon explication (p. 23-25 de la seconde édition [12]) de ce qui est requis d'une théorie de la vérité pour qu'elle

12. [N.D.T] Horwich, 1998b.

> soit adéquate n'inclut pas la conservativité, et (b) que je ne fais jamais allusion à la conservativité —et, en particulier, je ne suggère jamais que la théorie minimale est conservative. (Communication personnelle, citée dans (Tennant, 2010, p. 439))

À notre connaissance il s'agit là des seuls commentaires publiés de la part d'Horwich sur la question de la conservativité et de ses liens avec les thèses déflationnistes. Dans le contexte de la discussion de l'argument de Ketland et Shapiro, les déclarations d'Horwich peuvent surprendre. Il proclamme qu'il ne se considère pas soumis à une contrainte de conservativité tout en revendiquant à nouveau la théorie minimale. Mais, la théorie minimale, dans sa version typée possède justement de « bonnes » propriétés de conservativité. [13] Le problème, si problème il y a, serait plutôt que ces **T**-équivalences sont trop faibles pour satisfaire la contrainte de réflexivité. Or, à ce sujet Horwich ne dit mot.

À la fin de son article [14], Tennant considère d'ailleurs que, malgré les dénégations citées ci-dessus, « il est raisonnable d'interpréter les déclarations d'Horwich d'une manière qui le place [parmi les auteurs déflationnistes conservatifs] » [15], étant donné les difficultés d'interprétation que soulèvent les propos d'Horwich. Néanmoins, il faut rappeler que dans son ouvrage [16], Horwich donne une version non typée de sa théorie minimale, dans laquelle des formules [17] contenant déjà le prédicat de vérité sont autorisées à apparaître dans les **T**-équivalences . Or, cette « itération » du prédicat « Vr » peut produire des extensions non-conservatives, et même hautement non-conservatives puisqu' en l'absence de précision supplémentaire, elles peuvent produire des extensions incohérentes [18].

13. $T_d = \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi / \varphi \in \mathcal{L}_T\}$ est conservative sur la syntaxe, ou disons, sur une théorie contenant un minimum d'arithmétique.

14. Tennant, 2010.

15. Voir Tennant, 2010, tableau p. 447 et note de bas de page. Tennant prend néanmoins soin de noter « Horwich* » cet auteur déflationniste acceptant la contrainte de conservativité, pour le distinguer de ce qu'Horwich lui-même semble revendiquer.

16. Horwich, 1998b.

17. Horwich ne précise pas vraiment lesquelles. Voir cependant Horwich, 1998b, p. 40-42.

18. Nous faisons bien sûr ici référence au paradoxe du menteur et autres paradoxes sémantiques. La discussion de l'argument de Shapiro et Ketland, s'est appuyée sur des théories de la vérité typées, et dans ce cas, la distinction entre langage objet $\mathcal{L}_T$ et métalangage $\mathcal{L}' \supset \mathcal{L}_T \cup \{Vr\}$ permet de bloquer ces paradoxes. Horwich ne semble pas favorable à cette sorte de solution. Mais il reste très évasif sur la manière dont il faudrait traiter ces problèmes dans la perspective de sa théorie minimale. Aux pages 40 à 42 de son ouvrage, Horwich ne paraît pas prêt à renoncer à la logique classique, ni à adopter une solution tarskienne fondée sur la distinction langage-objet/métalangage. Il semble plutôt suggérer que les **T**-équivalences non typées soient expurgées d'assez de leurs instances pour ne plus mener aux paradoxes, mais en en conservant le plus possible de manière à conserver un système cohérent (voir Horwich, 1998b, p. 40-42 et 136). Malheureusement ce type de spécification est un peu court et soulève de nombreux problèmes techniques. Pour une argumentation détaillée sur le caractère un peu « léger » des

Peut-être est-ce dans cette perspective qu'il faut comprendre les propos d'Horwich.

Pour achever ce tour d'horizon des positions sur la question de la contrainte de conservativité, citons encore un auteur, proche du déflationnisme, qui s'est régulièrement prononcé contre cette contrainte. Dans une série d'articles consacrés à une étude des positions déflationnistes appuyée sur de nombreux résultats techniques, Volker Halbach, s'est régulièrement déclaré opposé à la contrainte conservativité (*cf.* Halbach, 2001b,c ; Halbach et Horsten, 2003). Selon lui, la thèse du déflationniste d'après laquelle le prédicat de vérité serait uniquement un instrument de généralisation et de « montée sémantique » ne le contraint nullement à être un outil peu puissant et ne pouvant avoir de conséquences « substantielles ».

Voici un passage représentatif de la position de Halbach :

> Le déflationniste doit donc accepter que sa théorie de la vérité a des conséquences mathématiques « substantielles » et qu'elle n'est en aucun cas conservative.
>
> Mais, là encore, le déflationniste ne devrait pas se sentir coupable. [...] La vérité ne sert d'autre but en dehors de celui d'exprimer et de prouver des généralisations ; mais un instrument de généralisation est un outil puissant, et qu'il ne serve pas d'autre but n'implique pas qu'il soit un outil sans tranchant. (Halbach, 2001b, p. 189) [19]

spécifications proposées par Horwich, si on les inteprète comme la recherche d'ensembles maximalement consistants de **T**-équivalences , voir McGee (1992). Malgré ces remarques, il faut reconnaître que la question des paradoxes est évidemment un immense problème pour toutes les analyses de la vérité, qu'elles soient d'inspiration déflationniste ou pas ; immense problème que nous avons délibérément et prudemment laissé de côté dans le cadre de ce travail.

19. Curieusement, dans un article plus ancien, Volker Halbach, propose une analyse des thèses déflationnistes dans laquelle ils les rapproche du programme de Hilbert et dans laquelle il invoque lui aussi une contrainte de conservativité :

> Ce fait *[N.D.T. : ici, Halbach fait référence à la conservativité de T_d sur T, c'est-à-dire à la conservativité d'une théorie purement décitationnelle* typée *($\{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi / \varphi \in \mathcal{L}_T\}$) sur une théorie arithmétique de base]* autorise une conception non-réaliste concernant le prédicat de vérité. [...] Ainsi, la notion décitationnelle de vérité peut être pensée en termes d'instrumentalisme, de formalisme, de théorie de la survenance et ainsi de suite. [...] **Toute théorie de la vérité qui contribue véritablement à la preuve d'énoncés « réels », c'est-à-dire d'énoncés du langage de base** *[N.D.T, selon notre terminologie : d'énoncés de $\mathcal{L}_T$]*, **n'est pas acceptable pour le décitationniste, puisqu'alors la vérité ne serait plus uniquement un outil pour exprimer des conjonctions infinies, mais un instrument indispensable dans des arguments à l'appui de thèses portant sur des faits non sémantiques.** La théorie des équivalences tarskiennes satisfait cette exigence de par sa conservativité sur sa théorie de base (Halbach, 1999b, p. 19-20, nous mettons en gras).

Il nous semble que ce type d'argumentation est on ne peut plus dans la droite ligne des arguments de

Ces dissonances illustrent sans doute une fois de plus la diffculté à définir précisément ce qu'est le déflationnisme. Chaque auteur paraît avoir sa propre compréhension du terme et de ses implications.

4.1.2 La position de Cieśliński (2017)

Dans un livre paru récemment, *The Epistemic Lightness of Truth : Deflationism and its Logic* (2017), Cezary Cieśliński a proposé une analyse systématique de la contrainte de conservativité à la lumière des débats qui ont agité les cercles déflationnistes depuis une vingtaine d'années. Pour développer sa propre conception déflationniste, il revient en détails sur ce qu'il appelle les motivations philosophiques de la contrainte de conservativité [20] et tente de déterminer si une telle contrainte s'impose de manière inévitable au déflationniste.

Cieśliński distingue soigneusement la contrainte de conservativité modèle-théorique [21] de la contrainte de conservativité syntaxique [22] car selon lui les motivations ou les arguments avancés à l'appui de l'une ou de l'autre de ces contraintes sont de natures différentes. Concernant la conservativité modèle-théorique, le diagnostique de Cieśliński est sans appel :

> Aucun argument convaincant n'a été présenté dans la littérature à l'appui de la thèse selon laquelle la conservativité [modèle-théorique] [23] découle de ce que les déflationnistes ont réellement dit. (Cieśliński, 2017, p. 172)

Dans son chapitre Cieśliński (2017, chapitre 9, p.145-156) examine plusieurs variantes argumentatives possibles en faveur de ce type de contraintes de conservativvité [24]. Selon

Shapiro et Ketland, et qu'ici, Halbach souscrit totalement à la contrainte de conservativité (le lien tracé avec le programme de Hilbert et les résultats techniques qui s'en sont suivis, nous semble d'ailleurs tout à fait pertinent à ce titre). Il semble donc qu'Halbach ait changé d'avis sur cette question. Depuis, il a néanmoins réitéré son refus de la contrainte de conservativité sans varier (voir, par exemple, Halbach (2014, 2001b,c) ; voir aussi la note de bas de page de Tennant (2010, p. 447), dans laquelle l'auteur indique que Volker Halbach lui a déclaré en privé qu'il ne souhaitait pas être classé parmi les déflationnistes conservatifs).

20. *Cf.* Cieśliński (2017, chapitre 9, p 145-713).

21. d'après laquelle tout modèle (*i.e.* structure d'interprétation) de la théorie de base doit pouvoir être enrichi en un modèle de la théorie étendue. *Cf.* 3.1.1 pour plus de détails.

22. d'après laquelle tout théorème exprimé dans le langage de la théorie de base et prouvable dans la théorie étendue doit déjà être prouvable dans la théorie de base. *Cf.* 3.1.1 pour plus de détails.

23. Cieśliński (2017) désigne ce type de propriété sous le nom de conservativité « sémantique ». Nous avons modifié la terminologie pour qu'elle soit cohérente avec la manière dont nous employons ces termes dans le cadre de ce travail.

24. Cieśliński (2017) se place dans le contexte d'extensions aléthiques d'une théorie de base arithmé-

lui, elles font toutes appel, explicitement ou implicitement, à la notion de modèle standard (de l'arithmétique) et à une identification de la vérité *simpliciter* à la vérité dans une structure d'interprétation :

> La difficulté fondamentale est que les arguments en faveur de la conservativité modèle-théorique semblent prendre pour acquise la notion de modèle standard ; c'est-à-dire que nous ne voulons pas que nos axiomes pour la vérité excluent des modèles étant donné que ces modèles sont désirables (ou attendus). (Cieśliński, 2017, p. 173)

Or, selon Cieśliński (2017), si les déflationnistes peuvent tout à fait se prévaloir des outils classiques de la théorie des modèles, notamment de la notion de vérité dans une structure [25], ils refuseront en revanche toute identification de la vérité « tout court » avec la vérité dans une structure :

> ce que [le déflationniste] ne peut pas faire, c'est présenter une description de la vérité arithmétique *simpliciter* comme la vérité dans un modèle choisi (standard ou attendu) de l'arithmétique. (Cieśliński, 2017, p. 147)

Bien au contraire, une thèse centrale du déflationnisme est que la notion de vérité (arithmétique)

> doit être caractérisée au moyen d'axiomes simples [...] qui jouent le rôle de postulats de signification. (Cieśliński, 2017, p. 146)

Dès lors, les arguments en faveur de la conservativité modèle-théorique qui s'appuient sur la notion de vérité dans un modèle attendu entrent en contradiction avec les thèses de base du déflationnisme :

> ce type d'argumentations en faveur de la conservativité modèle-théorique compromet la nature auto-suffisante de la caractérisation axiomatique de la vérité. (Cieśliński, 2017, p. 173)

---

tique et mentionne trois lignes d'argumentatives principales :

- une première variante s'appuie sur le fait que certaines extension non modèle-théoriquement conservatives peuvent exclure le modèle standard de l'arithmétique ;
- une seconde variante s'appuie sur l'identification de la vérité arithmétique à la vérité dans un modèle, sans préjuger de ce que sera un tel modèle, standard ou autre ;
- la troisième variante laisse apparemment de côté la notion de modèle standard (ou attendu) et entend placer tous les modèles de la théorie de base sur un pied d'égalité.

Voyez Cieśliński (2017, p. 145-156) pour plus de détails.

25. *Cf.* Cieśliński (2017, p. 147).

De tels arguments sont donc sans force d'un point de vue déflationniste [26].

Concernant le second type de contrainte de conservativité, à savoir la conservativité déductive (ou syntaxique), la position de Cieśliński (2017) est encore plus surprenante. Cieśliński déclare tout d'abord qu'à ses yeux,

> il n'existe qu'un seul candidat sérieux au rôle d'argument permettant d'établir une demande de conservativité à partir des anciennes doctrines déflationnistes. Cela consiste, en gros, à dériver la contrainte de conservativité syntaxique à partir de thèses instrumentalistes. En effet, certains déflationnistes ont soutenu que le concept de vérité n'était qu'un simple outil dont on devrait, en principe, pouvoir se passer dans les explications de faits non sémantiques ou dans les justifications de croyances non sémantiques. (Cieśliński, 2017, p. 157-158)

Ainsi à partir des thèses déflationnistes, d'inspiration instrumentaliste, affirmant l'absence de rôle explicatif ou justificatif de la notion de vérité, il doit être possible de déduire une contrainte de conservativité syntaxique. Cette fois, Cieśliński distingue les arguments s'appuyant sur l'absence de rôle explicatif de ceux partant de la notion de justification.

En ce qui concerne la possibilité de dériver une contrainte de conservativité syntaxique à partir de l'absence de rôle explicatif, Cieśliński déclare tout d'abord que :

> le principal obstacle qui rend difficile d'évaluer [ce type] d'arguments est que le concept d'explication en mathématiques n'est actuellement ni bien compris ni suffisamment étudié. (Cieśliński, 2017, p. 161)

S'appuyant sur les débats qui opposent encore aujourd'hui les philosophes des mathématiques, Cieśliński rappelle en effet que la notion d'explication en mathématiques reste très controversée et qu'on serait bien en peine de trouver ne serait-ce qu'une seule démonstration ou principe mathématiques sur lesquels tous s'accordent pour les considérer

26. Sans entreprendre d'analyse approfondie du raisonnement de Cieśliński (2017), contentons-nous de signaler ici que sa démonstration n'a pas entièrement convaincu Leon Horsten, autre auteur globalement favorable au déflationnisme. Dans sa recension du livre de Cieśliński, Horsten (2018) déclare :

> Cieśliński ne donne pas vraiment toute sa chance à la conservativité modèle-théorique. [... Et] un des arguments en faveur de la conservativité modèle-théorique discuté par Cieśliński semble plutôt prometteur.

Plus précisément, il s'agit de la troisième ligne argumentative examinée par Cieśliński, celle qui consiste à mettre sur un pied d'égalité tous les modèles de la théorie de base. Pour plus de détails sur ce débat qui s'annonce entre Horsten et Cieśliński voyez Cieśliński (2017) et Horsten (2018).

comme explicatifs [27]. Bien plus, à partir d'exemples choisis, Cieśliński tente de montrer qu'il existe en mathématiques des extensions non-conservatives qui ne fournissent pas de nouvelles preuves dans lesquelles les notions nouvellement introduites seraient investies de pouvoir explicatif. Et, à l'inverse, il présente également des cas où des notions nouvellement introduites, quoiqu'elles produisent des extensions conservatives et qu'on puisse donc en principe s'en passer, jettent néanmoins une lumière nouvelle sur des résultats déjà prouvés par ailleurs et en fournissent donc une nouvelle (voire meilleure) explication. S'il reconnaît lui-même que

> [la] notion d'explication dans les contextes mathématiques demeure obscure et [que] les exemples avancés ici peuvent être contestés (Cieśliński, 2017, p. 165)

Cieśliński en conclut néanmoins que

> l'équivalence « non-explicatif ≡ syntaxiquement conservatif » est problématique. La conservativité syntaxique *per se* ne garantit pas l'inexistence de preuves explicatives appuyées sur une théorie de la vérité, tandis que la non-conservativité syntaxique *per se* n'implique pas l'existence de telles preuves. (Cieśliński, 2017, p. 173)

Pour ce qui est de l'absence de poids justificatif de la vérité, Cieśliński argumente également contre l'idée que la non-conservativité d'une théorie aléthique serait incompatible avec le rôle non justificatif de la vérité déflationniste. Il se concentre cette fois sur les preuves sémantiques de cohérence d'une théorie arithmétique de base et souligne que la valeur justificative de telles preuves est notoirement problématique. En effet, si la théorie étendue au moyen d'axiomes pour la vérité permet bien (sous certaines hypothèses) de dériver la cohérence de la théorie de base [28], elle ne le fait qu'en s'appuyant justement sur certains principes de preuve déjà contenus dans la théorie d'origine [29]. Dans la mesure où les preuves formulées dans la théorie étendue emploient des principes de preuves plus forts (et donc plus « risqués ») que ceux de la théorie de base, la valeur justificative de telles preuves de cohérence ne semble guère probante. Ainsi, quelqu'un

27. Un principe aussi fondamental et élémentaire que l'induction sur les nombres entiers est ainsi loin de faire consensus (*cf.* Cieśliński (2017, p. 161-162)). Pour un aperçu global de l'état des débats sur la notion d'explication en mathématiques voyez Mancosu (2018).

28. Rappelons par exemple que l'extension (non-conservative) PA^{+ind}_{Tar} prouve $Con(PA)$. Voyez le chapitre 3 pour plus de détails.

29. Ainsi $PA^{+ind}_{Tar} \vdash Con(PA)$ en employant les axiomes de PA et l'induction non seulement pour $\mathcal{L}_{PA}$ mais aussi pour $\mathcal{L}_{PA \cup Vr}$.

qui douterait authentiquement de la fiabilité ou de la cohérence des principes de preuve contenu dans PA ne sera gère rassuré par une preuve de $Con(PA)$ appuyée sur des principes de preuves... de PA ! Cieśliński en conclut alors qu'

> il est parfaitement possible qu'une théorie de la vérité prouve de nouveaux théorèmes tout en ne procurant que fort peu, voire strictement en rien, pour ce qui est de leurs justifications. (Cieśliński, 2017, p. 169)

Au total, Cieśliński considère donc

> [les] explications [déflationnistes] traditionnelles ne permettent pas la dérivation de la contrainte de conservativité syntaxique en tant que caractéristique *obligatoire* des théories déflationnistes de la vérité. (Cieśliński, 2017, p. 170, nous soulignons)

Parvenu à ce point de son argumentation, Cieśliński opère un mouvement qui a de quoi surprendre. Quelques lignes après le passage que nous venons de citer, il poursuit en effet :

> Cependant, rien de ce qui a été dit ici n'implique que la conservativité syntaxique ne puisse fonctionner comme une *nouvelle* explication de la légèreté de la vérité, proposée en étant pleinement conscient que son lien avec la tradition est assez lâche. (Cieśliński, 2017, p. 170, italiques de l'auteur)

De même lorsqu'il résume les conclusions de son chapitre consacré aux motivations philosophiques de la conservativité, Cieśliński déclare :

> toutefois, il vaut la peine de considérer la conservativité syntaxique comme une *nouvelle* explication de la légèreté de la vérité, réliée seulement de manière relâchée à la tradition philosophique. (Cieśliński, 2017, p. 173)

En résumé, après avoir longuement argumenté contre l'idée qu'une contrainte de conservativité, modèle-théorique ou syntaxique, découle des thèses déflationnistes traditionnelles et de ce que les déflationnistes ont « réellement dit », Cieśliński (2017) finit par réintroduire la conservativité syntaxique comme caractéristique essentielle de sa propre position déflationniste. Il la présente en effet comme une explication technique précise de la thèse informelle de la légèreté épistémique de la vérité, thèse qu'il place au coeur du déflationnisme tout au long de son ouvrage [30]. Cette position est pour le moins

30. comme l'atteste du reste le titre de son livre *The Epistemic Lightness of Truth : Deflationism and its Logic*.

étonnante. On peut la considérer comme le signe d'une grande probité intellectuelle : alors même qu'il entend caractériser sa propre position déflationniste au moyen de la conservativité syntaxique, Cieśliński reconnaît que cette caractérisation ne découle par directement des thèses déflationnistes formulées par ses prédécesseurs. Mais on peut également soupçonner qu'il y a là une tension potentielle au sein de l'analyse de Cieśliński (2017). Comment expliquer que la conservativité syntaxique soit une caractéristique essentielle du déflationnisme « à la Cieśliński » alors même que, selon lui, elle ne résulte nullement des thèses déflationnistes ? Par exemple, si l'on fait de la « légèreté épistémique » de la vérité une thèse informelle centrale au déflationnisme en général —ancien ou moderne, traditionnel ou « à la Cieśliński »— comment expliquer que cette thèse, assez vague, se traduise de manière précise par une contrainte de conservativité dans la cas du déflationnisme « cieślińskien » tandis qu'elle ne débouche nullement sur une telle contrainte dans le cas des déflationnistes classiques ? Présenter cette caractérisation comme une nouvelle explication « reliée seulement de manière relâchée à la tradition » suffit-il à lever les doutes ? Nous laisserons cette question en suspend ici et remettons une analyse plus poussée de cet aspect du travail de Cieśliński (2017) à un travail ultérieur.

4.1.3 Pouvoir expressif et conservativité

À titre de bilan provisoire, nous pouvons néanmoins dire ceci : à première vue, la conservativité (syntaxique) nous semble une traduction technique précise tout à fait naturelle des thèses déflationnistes à propos de la vérité ; essentiellement pour les raisons déjà invoquées, à savoir que si la vérité n'est pas une propriété « métaphysiquement substantielle », si le prédicat « vrai » n'est qu'un outil logico-syntaxique dénué de pouvoir explicatif, alors on ne voit pas comment le fait d'asserter ou de supposer la vérité d'une théorie objet (en plus de la théorie elle-même) pourrait nous permettre d'établir de nouveaux faits non sémantiques, comment le fait de parler en termes de vérité pourrait augmenter notre capacité à prouver des faits qui ressortissent du domaine de discours de notre théorie de base, que ce soit les entiers naturels, les atomes, ou ce qu'on voudra. Toutefois, au vu des controverses suscitées par ce critère et à la lumière de l'analyse proposée par Cieśliński (2017), il nous faut bien reconnaître qu'il serait exagéré de dire qu'une contrainte de conservativité s'impose de manière irréfragable ou obligatoire [31] aux déflationnistes. Il y a plusieurs raisons à cela. Tout d'abord, il est vrai que la contrainte

31. selon les propres termes de Cieśliński (2017, p. 170).

de conservativité, sous une forme ou sous une autre, a été introduite par les détracteurs du déflationnisme afin de l'employer contre lui. Rares sont donc les auteurs déflationnistes a avoir ouvertement accepté une contrainte de conservativité, Cieśliński (2017) consitiuant ici l'exception paradoxale qui confirme la règle[32]. De plus et de manière plus importante, il faut se souvenir que les thèses déflationnistes sont souvent avancées de manière informelle et intuitive, prenant parfois la forme de quasi slogans. Dès lors, la traduction de ces thèses sous la forme d'un critère précis permettant leur évaluation laisse une large part à l'interprétation et donc au débat exégétique : l'articulation entre les déclarations philosophiques générales comme « la vérité n'est pas une propriété métaphysiquement robuste » ou « le prédicat de vérité n'est qu'un outil de décitation sans force explicative » et les outils formels qui permettent d'obtenir des résultats précis est toujours délicate. En somme, les thèses déflationnistes sont (peut-être) trop vagues ou trop floues pour pouvoir être strictement identifiées à un critère technique précis, quel qu'il soit[33]. Mais quoi qu'il en soit et quoi que l'on pense des liens plus ou moins étroits entre les thèses déflationnistes et la conservativité, force est de constater que la contrainte de conservativité est omniprésente dans toute la littérature techniquement informée portant sur le déflationnisme aléthique. Après plus de vingt ans de controverses et de discussions, la contrainte de conservativité, dans l'une ou l'autre de ses variantes, est à notre connaissance le seul critère technique rigoureux qui ait été proposé comme traduction ou explication précise des thèses déflationnistes en matière de vérité. C'est au point que même ceux qui le considèrent comme un critère imparfait ou mauvais finissent par l'adopter ou en tout cas par lui faire une place centrale au sein de leurs discussions. Nous nous conformerons ici à cet usage.

Pour poursuivre quelque peu notre analyse, il est important de garder en mémoire que les déflationnistes contemporains ne sont pas partisans d'une théorie éliminativiste ou théorie de la redondance vis-à-vis de la vérité. Bien au contraire, ils reconnaissent

32. Mais voyez également le début de ce chapitre, notamment les positions de Tennant (2002, 2010) et celle de Field (1999).

33. Un cas limite illustrant ce type de problèmes inspiré par certaines analyses de Cieśliński (2017) :

1) Le déflationnisme se caractérise par le fait qu'il considère la vérité comme dénuée de toute capacité explicative.

2) Cette absence de capacité explicative doit-elle se traduire par une propriété de conservativité ?

3) On ne peut répondre à cette question de manière ferme et définitive dans la mesure où la notion d'explication n'est pas suffisamment comprise.

Nous voilà bien avancés ! Mais si la notion d'explication n'est pas suffisamment comprise, qu'en est-il alors du déflationnisme en tant que théorie affirmant l'absence de pouvoir explicatif de la vérité ?

pleinement et insistent sur le caractère indispensable du prédicat de vérité. Mais, nous disent-ils, cette indispensabilité n'est pas due au fait que la vérité serait une propriété substantielle et importante dont la possession ou non par telle ou telle entité pourrait entrer en ligne de compte dans l'explication de tel ou tel phénomène. Elle est simplement due à des contraintes de capacités expressives et au fait qu'un prédicat décitationnel est nécessaire pour exprimer certaines généralisations. Une telle conception laisse apparemment ouverte la possibilité que la notion de vérité *puisse apparaître* dans nos explications, ou disons dans nos théories permettant d'établir tel ou tel résultat, sans pour autant être investie de rôle explicatif, qu'elles y apparaissent comme un simple mais *indispensable* « auxiliaire expressif ». Il faudrait donc distinguer et séparer indispensabilité expressive et pouvoir explicatif. Ces considérations offrent peut-être une voie de réponse pour le déflationnisme face au problème de la non-conservativité.

Cette ligne d'argumentation soulève toutefois un redoutable problème méthodologique. Après tout, au moins à première vue, n'importe quelle expression nouvelle que l'on peut introduire dans notre langage augmente en un sens les capacités expressives de ce dernier.[34] En ajoutant à un langage qui en serait dépourvu les terme « rouge », « électron », ou « avoir une charge électrique négative »,[35] j'augmente les capacités expressives de ce langage. Je peux à présent exprimer des choses comme : « cet objet x est rouge », « ce morceau de brique rouge est composé d'électrons », « tout électron possède une charge électrique négative », *etc.* De même, en ajoutant à un langage de la physique

34. Du moins si cette nouvelle expression n'y est pas déjà définissable. Dans le cas contraire, la nouvelle expression introduite n'augmente en rien les capacités expressives du langage puisqu'on pourra toujours remplacer la nouvelle expression par sa définition. Nous restons délibérément vagues et peu précis sur la question de savoir quand et à quelles conditions une nouvelle expression ajoutée à un langage en augmente la puissance expressive. En effet, ce n'est pas ce point qui pose problème pour la discussion du déflationnisme, dans la mesure où toutes les parties s'accordent (à peu près) sur le fait que le prédicat de vérité augmente les capacités expressives de notre langage. La question est de savoir s'il ne fait que cela. Néanmoins, signalons que dans leur ouvrage consacré au nominalisme, Burgess et Rosen discutent ce problème de déterminer à quelles conditions l'extension d'un langage et d'une théorie est expressivement plus riche que le langage et la théorie étendus. Pour ce faire, ils introduisent la notion d'*extension expressivement conservative* : soit T' exprimée dans un langage $\mathcal{L}'$ sorté avec deux sortes de variables : $\vec{x}$ et $\vec{\mathrm{x}}$, soit $T \subseteq T'$ la restriction de T' au langage unisorté $\mathcal{L}$ qui ne contient que des variables de la sorte $\vec{x}$. T a pour axiomes les axiomes de T' qui sont exprimés dans le seul $\mathcal{L}$ (les axiomes contenant du vocabulaire (prédicats ou variables $\vec{\mathrm{x}}$) pris dans $\mathcal{L}'/\mathcal{L}$ sont laissés de côté). Alors T' est une extension expressivement conservative de T ssi pour toute formule $\varphi'(\vec{x})$ du langage $\mathcal{L}'$ dont les seules variables libres sont de la première sorte, il existe une formule $\varphi(\vec{x})$ de $\mathcal{L}$ ayant les mêmes variables libres telle que : $T' \vdash \forall\vec{x}(\varphi'(\vec{x}) \leftrightarrow \varphi(\vec{x}))$ (voir Burgess et Rosen, 1997, § 1.B.2.b). Comme on pouvait s'y attendre, *mutatis mutandis*, il est clair que les extensions aléthiques considérées ici (T_d, T_{Tar}, *etc.*) ne sont pas expressivement conservatives en ce sens.

35. et à supposer que ces expressions n'étaient pas déjà définissables dans le langage d'origine.

l'expression « licorne », j'en accrois la force expressive. Je puis à présent affirmer des choses comme : « il y a une licorne au centre du soleil » ou bien « cette licorne est constituée d'électrons chargés négativement ». Dès lors, lorsqu'on enrichit un langage et une théorie donnés en leur ajoutant un nouveau terme, éventuellement muni de nouveaux axiomes permettant d'en gouverner l'emploi, et que ce faisant on augmente nos capacités expressives, comment déterminer les cas où l'expression nouvelle n'est qu'un outil purement expressif sans contenu nouveau ni capacité explicative supplémentaire, et les cas où cette expression désigne une notion véritablement importante et jouant un rôle explicatif ? Le point crucial est évidemment de déterminer si une théorie formalisant notre emploi d'un concept sans contenu substantiel et dénué de pouvoir explicatif, qu'il soit ou non *par ailleurs* expressivement indispensable, peut néanmoins être non-conservative.

Si l'on en juge par les réflexions des philosophes et des scientifiques sur ce type de questions *précédant* le débat sur le déflationnisme, on serait tenté de répondre qu'évidemment non. À nos yeux, il n'est pas exagéré de dire qu'historiquement le critère de conservativité a précisément été introduit pour saisir l'absence de contenu substantiel et de rôle explicatif dévolu à telle ou telle notion, tels ou tels ensemble de principes ou champs théoriques. Nous avons rappelé dans le chapitre précédent deux exemples paradigmatiques de ce type de démarche. Incontestablement, pour Hilbert un résultat de conservativité était censé garantir et justifier l'absence de contenu réel des mathématiques idéales tout comme leur inanité explicative. *A contrario*, un résultat de non-conservativité semble établir que les axiomes des mathématiques transfinies ont bel un bien un contenu et que ces notions tiennent bel et bien un rôle dans les démonstrations de certains de nos théorèmes relevant des mathématiques finitistes. [36] De même, selon la défense fieldienne du nominalisme développée dans *Science Without Numbers*, ce qui distingue les fictions mathématiques, sans contenu réel ou substantiel et dénuées de véritable pouvoir explicatif, des entités théoriques inobservables de la physique qui, elles, sont substantielles et jouent certainement un rôle important dans nos explications, c'est la conservativité des unes et la non-conservativité des autres. [37] Autrement dit, dans cet

36. Aujourd'hui encore, la recherche de résultats de conservativité est centrale dans les programmes néo-hilbertiens contemporains (qui donc doivent faire avec les théorèmes d'incomplétude de Gödel). Nous pensons ici aux mathématiques inversées (reverse mathematics), au finitisme ou à l'ultra-finitisme. Pour une discussion récente de l'emploi de la notion de conservativité dans une perspective finitiste concernant les mathématiques, voir par exemple Burgess (2010).

37. respectivement sur la physique nominalisée $\mathcal{PN}$, et sur la physique des observables $\mathcal{PO}$. Voyez le chapitre précédent pour plus de détails. Rappelons simplement ici le passage crucial de *Science Without Numbers* :

ouvrage, Field fait de la non-conservativité le critère par excellence de notre engagement ontologique vis-à-vis de la nature substantielle et de la fertilité explicative des notions apparaissant dans nos constructions théoriques.

Sans doute n'existe-t-il pas en philosophie de principe méthodologique intangible. Et peut-être que le lien tracé entre susbstantialité et pouvoir explicatif d'une part, et non-conservativité d'autre part, tel qu'il ressort des analyses portant sur d'autres concepts que la vérité, ne peut s'appliquer tel quel lorsqu'on examine les thèses déflationnistes en matière de vérité. Mais on est aussi en droit de se demander pourquoi : pourquoi devrions-nous renoncer ou mettre de côté un principe méthodologique (relativement) bien ancré lorsqu'il s'agit d'évaluer la plausibilité de certaines thèses déflationnistes ? Pourquoi la vérité déflationniste serait-elle l'exception qui confirme la règle ? Face à un tel problème, il est clair qu'on ne peut se limiter à déterminer si les auteurs déflationnistes ont explicitement ou implicitement adopté ou récusé une contrainte de conservativité. De même, il n'est guère plus satisfaisant de se contenter de dire, sans plus d'argument, que les déflationnistes considèrent le prédicat de vérité comme un « outil purement expressif » mais que ce caractère d'outil purement expressif ne contraint nullement la vérité à n'avoir aucune conséquence substantielle ou à être dénuée de pouvoir explicatif. Car que faut-il comprendre par outil « purement expressif » si ce n'est une manière de sous-entendre que l'outil en question n'est qu'expressif au sens où il ne désigne pas une propriété substantielle et où il n'a pas de rôle explicatif à jouer au sein de nos théories ?

> *Mais il existe une différence fondamentale entre ces deux cas* [N.D.T. Field fait ici références au cas des entités inobservables de la physique, supposées exister réellement, par opposition à celui des entités fictives composant les mathématiques]*, et cette différence réside dans la nature des lois-ponts.* Dans le cas des particules subatomiques, la théorie $\mathcal{T}$, à présent interprétée de manière à inclure les lois-ponts (ainsi que, peut-être, quelques hypothèses concernant les conditions initiales), peut être appliquée à un ensemble de prémisses portant sur les observables d'une manière à engendrer des assertions authentiquement nouvelles à propos des observables, assertions qui ne seraient pas dérivables sans l'aide de $\mathcal{T}$. Mais, dans le cas des mathématiques, la situation est tout à fait différente : ici, si nous prenons une théorie mathématique qui comprend les lois-ponts (*i.e.* qui comprend des énoncés assertant l'existence de fonctions reliant les objets physiques à certains objets abstraits "purs", et peut-être notamment des assertions obtenues au moyen d'un principe de compréhension qui emploie en même temps et du vocabulaire physique, et du vocabulaire mathématique), alors les mathématiques sont applicables au monde, *i.e.* elles sont utiles en ce qu'elles nous permettent de tirer des conclusions nominalistiquement formulables à partir de prémisses nominalistiquement formulables ; *mais là, contrairement au cas de la physique, les conclusions auxquelles nous parvenons par cet intermédiaire ne sont pas authentiquement nouvelles, elles sont déjà dérivables d'une façon plus ardue à partir des prémisses, sans recours aux entités mathématiques.* (Field, 1980, p. 10–11, italiques de l'auteur)

S'ils refusent l'emploi de la conservativité, c'est sans doute aux déflationnistes que revient la tâche d'expliquer plus précisément ce qu'ils entendent exactement par prédicat « purement expressif », par propriété « non-substantielle », ou par propriété « non-explicative ». Charge à eux également de fournir un critère technique précis permettant de trancher les cas litigieux autrement que par une pétition de principe ou un appel à l'intuition dont l'éventail des positions exposées à la section 4.1.1 montre bien qu'elle est loin d'être la chose du monde la mieux partagée, y compris au sein des rangs déflationnistes. À cette double tâche s'ajoute aussi celle de donner des raisons suffisantes et non *ad hoc* justifiant l'abandon du critère de conservativité traditionnellement utilisé jusque là pour trancher ce type de questions. À notre connaissance, une telle entreprise, si tant est qu'elle soit possible, n'a jamais été menée à bien. Mais cela ne doit pas nous empêcher de poursuivre notre analyse de l'argument de Ketland et Shapiro, et d'ailleurs de nombreux et très intéressants éléments de réponses partiels ont été avancés au cours du débat.

4.1.4 *Bis repetita* : conservativité et notions logiques

Parmi ces éléments de réponse partiels, l'un des arguments contre la contrainte de conservativité qui ne se limite à une fin de non recevoir a été proposé par Henri Galinon. [38] L'idée centrale est de montrer qu'en dépit de ce qui ressort de l'analyse et de l'emploi de la conservativité dans le cadre des programmes réductionnistes à la Hilbert ou à la Field, les résultats de non-conservativité sont malgré tout compatibles avec l'absence de rôle explicatif et le caractère purement expressif d'une notion. Pour ce faire, Henri Galinon invoque des résultats de non-conservativité qui concernent des notions qu'intuitivement nous serions tentés de considérer comme « purement expressives », « non explicatives » et « non substantielles », à savoir les constantes logiques. C'est en effet un fait bien connu, et notoirement problématique pour l'analyse de la logicité, que dans certaines circonstances des notions traditionnellement regardées comme logiques peuvent donner naissance à des extensions non-conservatives. Le cas le plus célèbre est sans doute celui de la non-conservativité de la logique classique sur la logique intuitionniste. Par exemple, en ajoutant à un système pour la logique intuitionniste (**LI**) un principe tel que le tiers-exclu, ou le raisonnement par l'absurde, on obtient une extension non-conservative

38. Originellement dans Galinon (2010), partiellement repris dans Galinon (2012), puis dans Galinon (2015).

(**LC**). Dans le système ainsi étendu on peut par exemple dériver les lois de Peirce : $((p \to q) \to p) \to p$, qui ne sont pas valides en logique intuitionniste.[39] Pour autant, les notions logique sont souvent avancées à titre d'exemples paradigmatiques de notions purement expressives servant uniquement à articuler des contenus donnés par ailleurs. Si de telles notions peuvent produire des extensions non-conservatives, peut-être la situation avec la vérité est-elle la même ; peut-être que d'éventuels résultats de non-conservativité ne remettent pas en cause le caractère purement expressif, non explicatif et sans contenu substantiel du prédicat « vrai ».

De manière plus détaillée, l'argument pourrait se formuler comme suit[40] :

1. Les notions logiques peuvent donner naissance à des extensions non-conservatives.
2. Les notions logiques sont intuitivement considérées comme des outils purement expressifs (respectivement des notions non-substantielles, resp. des notions dénuées de pouvoir explicatif propre).
3. De 1. et 2., on conclut que les résultats de non-conservativité concernant une notion N sont compatibles avec le caractère purement expressif de cette notion (respectivement avec sa non-substantialité, resp. son absence de pouvoir explicatif).

Remarquons que 3. enfonce déjà un coin dans la première prémisse du raisonnement anti-déflationniste développé par Ketland et Shapiro. Si la non-conservativité n'est pas (toujours) synonyme de substantialité, de capacité explicative et de nature non-purement expressive, alors peut-être que le déflationniste n'a pas à se soucier de savoir si sa théorie de la vérité possède de « bonnes » propriétés de conservativité.

Mais on peut encore renforcer cet argument en notant que :

4. La vérité, telle qu'elle est comprise par les déflationnistes, est (semblable à) une notion logique.

CL. De 4. on déduit que 3. s'applique parfaitement au cas de la vérité déflationniste

39. Voir le chapitre précédent 2, notamment l'annexe, pour plus de détails techniques et pour d'autres exemples.

40. *Avertissement important au lecteur* : l'argumentation de Galinon (2010) contre la contrainte de conservativité et en faveur du caractère purement expressif et non-explicatif de la notion de vérité ne se limite pas à l'évocation pour comparaison des cas de non-conservativité impliquant des constantes logiques. Cette comparaison ne constitue en quelque sorte que la première étape de l'argumentation (la comparaison avec le cas des constantes logiques se trouve p. 149-155, et les exemples tirés des constantes logiques se trouvent précisément p. 152-155 Galinon (2010)). L'argumentation se poursuit ensuite par une analyse proche de celle de Field (1999) portant sur le rôle que joue l'extension du schéma d'induction dans l'apparition de la non-conservativité (sur l'arithmétique de Peano) de l'extension aléthique PA^{+Ind}_{Tar}. Nous revenons sur cette partie de l'argumentation lorsque nous traitons la réponse de Field. *cf.* 4.2.

et que par conséquent d'éventuels résultats de non-conservativité ne remettent pas en cause le caractère purement expressif de cette notion (respectivement sa non-substantialité, resp. son absence de pouvoir explicatif). En d'autres termes, la (non)-conservativité n'est pas un problème pour le déflationnisme et la première contrainte avancée par Shapiro et Ketland n'est pas justifiée.

Il y a plusieurs manières de défendre la contrainte de conservativité contre l'argument ci-dessus.

La première remarque que nous formulerons se borne à rappeler que nous sommes en désaccord avec l'assertion 4 ci-dessus, c'est-à-dire avec la thèse selon laquelle la vérité est une notion logique. Comme nous l'avons déjà expliqué, [41] les justifications apportées jusqu'à présent à l'appui de cette thèse fréquemment attribuée aux déflationnistes ne nous semblent pas probantes. La prémisse 4. n'était pas totalement sans importance, dans la mesure où elle renforce l'argument en donnant l'impression (ou l'illusion) qu'on traite uniformément comme non-problématiques des cas de non-conservativité portant sur une même famille de notions (à savoir les notions logiques ou quasi-logiques auxquelles se rattacherait la vérité). Ceci étant, même si on rejette 4., l'argument conserve une certaine force. Après tout, les notions purement expressives (resp. non explicatives, resp. non substantielles) forment peut-être un large ensemble, dont les notions logiques ne seraient qu'une sous-partie propre. Le fait que la vérité ne soit pas une notion logique ne l'empêche pas d'appartenir à cet ensemble plus large. Ce qui compte pour l'argument, et ce que montre (ou semble montrer) (1. 2. et 3.), à travers le cas particulier des notions logiques, c'est que la non-conservativité ne remet pas en cause l'appartenance à cet ensemble plus large de notions purement expressives. Que la vérité soit logique ou non, la non-conservativité ne l'empêcherait donc pas d'être purement expressive.

Une autre façon de répondre à cet argument consiste à remettre en cause l'inférence ($\{1., 2.\} \vdash 3.$). Peut-être que l'enseignement à tirer de la non-conservativité de certains principes gouvernant des notions logiques est justement que nos intuitions de départ sont trompeuses : les notions logique ne sont pas si purement expressives ou non explicatives que l'on pouvait le penser. Reprenons l'exemple de la non-conservativité de la logique classique sur la logique intuitionniste. Comme nous venons de le rappeler, si on enrichit un système de déduction naturelle (**LI**) pour la logique intuitionniste au moyen d'un principe suffisamment fort pour obtenir la logique classique (**LC**), [42] la loi de Peirce

41. Pour plus de détails sur cette question, voyez le chapitre 2.
42. ce peut-être la principe du tiers-exclu, ou le raisonnement par l'absurde, ou bien encore la loi

devient dérivable, alors qu'elle n'est pas intuitionnistiquement valide. On a donc bien un cas de non-conservativité, et les notions entrant en jeu dans ce cas sont bien des notions logiques. [43] Mais quelle leçon devons-nous tirer de ce résultat ? Montre-t-il bien comme le suggère Henri Galinon que des notions purement expressives peuvent donner naissance à des extensions non-conservatives ? Ne faudrait-il pas au contraire y voir l'occasion de réviser la prémisse 2. ? On peut en effet s'interroger : dans cette dérivation de la loi de Peirce, dérivation qui vaut d'ailleurs justification aux yeux d'un logicien classique, quelles notions jouent un rôle explicatif ? Quelles hypothèses substantielles nous permettent de parvenir au résultat ? Sans doute ne s'agit-il pas des seules règles gouvernant les constantes logiques intuitionnistes (puisque celles-ci sont insuffisantes pour obtenir la loi de Peirce), sans doute le principe supplémentaire que nous avons ajouté pour obtenir (**LC**) est-il crucial. Mais quoi qu'il en soit, il semble difficile d'exonérer totalement les notions logiques (classiques) et les principes (classiques) les gouvernant de tout rôle explicatif dans cette preuve, puisque ce sont les seules notions et les seuls principes qui y apparaissent ! [44]

Dans la même veine, on pourrait défendre l'idée qu'accepter un principe comme le tiers exclu ou le raisonnement par l'absurde, c'est se donner un principe explicatif extrêmement fort. [45] En mathématique, on sait par exemple que ces principes sanctionnent les démonstrations non-constructives. Pour en rester au cas de l'arithmétique, ce qui distingue l'arithmétique classique de Peano (**PA**) de son pendant intuitionniste, parfois intitulée arithmétique de Heyting (**HA**), ce n'est pas l'adoption d'axiomes arithmétiques

d'élimination des doubles négations. Il y a de multiples possibilités équivalentes au sens où elles étendent toutes (**LI**) en un système pour la logique classique (**LC**).

43. Du moins selon l'acception classique du terme.

44. Sauf à considérer qu'il n'y a rien à expliquer dans le cas précis de la dérivation d'une loi logique (classique) telle que la loi de Peirce. Peut-être en effet que la loi de Peirce est vraie « par définition » et qu'on aurait tort d'attendre ici une quelconque explication. Mais si tel est le cas, on pourrait attendre quelques... explications de ce dernier point, ainsi qu'une clarification de la notion d'explication que le déflationniste a en vue. Et si c'est cette voie qu'entend emprunter un défenseur du déflationnisme, on pourra lui rétorquer qu'il est alors difficile de maintenir le parallèle entre ce résultat de non-conservativité portant sur la dérivation d'une loi logique et le cas de la non-conservativité de l'extension aléthique tarskienne T_{Tar}^{+Ind} qui, elle, débouche sur la preuve d'un théorème portant sur les nombres entiers (et même sur plusieurs preuves de plusieurs théorèmes portant sur les nombres entiers), lequel théorème réclame certainement une explication.

45. Nous ne pouvons pas ne pas citer le mot célèbre de Hilbert à ce sujet :

> « Oter le [*tertium non datur*] au mathématicien serait comme si on voulait enlever à l'astronome son télescope, au boxeur le droit de se servir de son poing. Interdire les théorèmes d'existence et proscrire le [*tertium non datur*] revient quasiment à renoncer à la science mathématique. » Hilbert (1927, p. 159, traduction de Jean Largeault)

supplémentaires : les axiomes arithmétiques des deux théories sont rigoureusement identiques [46] ; au contraire, c'est l'adoption de principes [47] logiques supplémentaires. Pour autant les deux théories ont certainement des contenus très différents. Et lorsqu'on examine une preuve dans (**PA**) valide d'un point de vue classique mais inacceptable pour un intuitionniste, c'est-à-dire lorsqu'on est précisément confronté à un cas de non-conservativité de (**PA**) sur (**HA**), peut-on vraiment soutenir l'idée que les notions logiques [48] ne jouent pas ici de rôle explicatif alors même que la seule chose qui distingue cette preuve d'une démonstration intuitionniste c'est l'emploi de notions et de principes logiques classiques ? Il nous semble que c'est loin d'être évident. En tout cas, il ne manque pas de logiciens, de mathématiciens et de philosophes pour considérer que des notions traditionnellement (ou plutôt classiquement) considérées comme logiques ont un contenu suffisamment substantiel et une puissance explicative suffisamment forte pour qu'ils refusent de les employer, ou à tout le moins pour qu'ils refusent de les considérer comme logiques. Bien entendu, il ne s'agit pas de reprendre ici à nouveaux frais la querelle opposant intuitionnistes et classiques sur la question de savoir quelle est la « bonne » logique. Nous tenons à rester sufffisamment et prudemment neutres sur cette question. Les remarques précédentes ont seulement pour but de jeter un doute raisonnable sur l'absence de rôle explicatif des notions logiques impliquées dans des cas de non-consertativité, c'est-à-dire sur l'absence de rôle explicatif joué par les notions logiques lorsque celles-ci nous donnent accès à des moyens de preuve authentiquement nouveaux. Pour être précis jusqu'au bout, disons que face à l'argument ($\{1., 2.\} \vdash 3.$), un intuitionniste à la Dummett rejettera la prémisse 1., ce qui l'amènera à refuser de considérer comme logiques certains principes classiques [49] À l'inverse, le contre-argument face à l'inférence ($\{1., 2.\} \vdash 3.$) que nous esquissons ici consisterait plutôt à accepter 1. mais à remettre en cause 2. Notez que dans les deux cas néanmoins l'argument en faveur de 3. est bloqué. [50]

46. Même s'ils ne sont peut-être pas lus exactement de la même manière selon que l'on a à faire à un mathématicien classique ou à un mathématicien intuitionniste.

47. *i.e.* de règles d'inférence, ou d'axiomes, selon le type de système de preuves choisi : méthode axiomatique à la Hilbert, déduction naturelle, *etc.* .

48. ou peut-être le choix d'interpréter ces notions de telle ou telle manière : interprétation classique *vs* intuitionniste de la négation, ou de l'implication, ou de la disjonction, *etc.* .

49. les propriétés de non-conservativité des lois de la logique classique sont au demeurant l'une des raisons principales invoquées par Dummett à l'appui de sa position (voyez le chapitre 2 pour plus de détails et des références sur cette question).

50. Encore une remarque d'ordre historique : lorsqu'il élaborait son fameux programme, l'un des problèmes auxquels était confronté Hilbert était l'instabilité de la position finitiste par l'application des lois de la logique classique. Ainsi la négation d'un énoncé finitiste n'est pas toujours un énoncé finitiste (voyez le chapitre précédent pour plus de détails 3.1.2). Malgré tout, Hilbert n'entendait pas

Si toutefois ces considérations ne suffisaient pas, signalons que l'un des arguments les plus radicaux *en faveur* de la force explicative des notions logiques, et donc en faveur de leur nature *non* purement expressive, a été employé dans un texte portant lui aussi sur une analyse du déflationnisme. Dans un article paru en 2005, [51] Nic Damnjanovic examine un autre argument fameux avancé contre la conception déflationniste de la vérité et habituellement baptisé « argument du succès ». Pour le dire rapidement, l' « argument du succès » contre le déflationnisme tente de montrer que la vérité joue bel et bien un rôle explicatif important dans l'explication du fait qu'un agent cognitif possèdant des croyances vraies verra ses chances de succès augmenter. [52] Le contexte de cette discussion est donc assez différent de celui qui nous occupe ici puisqu'il concerne le rôle de la vérité dans certaines explications causales, et il n'est pas question de conservativité dans l'article de Damnjanovic. Toutefois, les thèses déflationnistes sur la logicité de la vérité et sur son absence de pouvoir explicatif se trouvent là encore au centre de la discussion. Et le point qui nous intéresse est l'analyse que propose Damnjanovic du rôle explicatif des notions logiques en général et de la vérité déflationniste en particulier. Cette analyse va frontalement à l'encontre de l'idée que les notions logiques sont purement expressives

renoncer aux lois de la logique classique qui lui paraissaient indispensables en pratique et pour des raisons de clôture expressives du langage de la science. Un résultat de conservativité pour les mathématiques idéales était donc censé garantir que le détour hors des méthodes et des énoncés strictement finitistes, et notamment l'emploi de la logique classique, ne nous conduisait pas à adopter des principes explicatifs trop forts ou à prendre des engagements ontologiques dépassant ceux acceptables du point de vue finitiste. À travers la conservativité, il s'agissait donc bien de s'assurer, entre autre, de l'absence de pouvoir explicatif (supplémentaire par rapport aux principes finitistes) des notions de la logique classique. *Dans une perspective hilbertienne (ou post-hilbertienne)*, il est donc pour le moins curieux d'invoquer des résultats de non-conservativité concernant des notions logiques pour appuyer l'idée que des notions non-explicatives peuvent donner naissance à des extensions non-conservatives.

51. Damnjanovic (2005).

52. L'argument du succès a semble-t-il été suggéré pour la première fois dans Putnam (1978, chapitres I et II). On en trouve également une version dans Field (1986). Il en existe diverses variantes qui ont donné naissance à d'importantes discussions pour le déflationnisme en matière vérité. L'analyse de cet argument déborde le cadre du présent travail et nous n'en tenterons pas l'étude détaillée.

Voici la structure globale de l'argument d'après Damnjanovic (2005, p. 54-55) :

1. Si A a des croyances vraies à propos de la manière d'obtenir ce qu'il veut, A est plus susceptible d'obtenir ce qu'il veut. [platitude]
2. Par conséquent, si A a des croyances à propos de la manière d'obtenir ce qu'il veut qui possèdent la propriété d'être vraies, A est plus susceptible d'obtenir ce qu'il veut. [par 1. et par transformation pléonasitique]
3. Par conséquent, la propriété d'être vrai est employée dans une généralisation causalement explicative. [par 2. et par définition de généralisation causalement explicative]
4. Par conséquent, la vérité est une propriété causalement explicative. [par 3.]
5. Par conséquent, le déflationnisme est faux [par 4. et par définition du déflationnisme]

et dénuées de pouvoir explicatif ; autrement dit, elle fournit une raison supplémentaire de rejeter la prémisse 2. dans l'inférence ($\{1., 2.\} \vdash 3.$).

Lorsqu'il examine ce qu'il appelle la réponse standard des déflationnistes à l'argument du succès, Nic Damnjanovic distingue deux étapes [53] : la première consiste pour le déflationniste à montrer que l'emploi de la notion de vérité dans l'explication du succès d'un agent possédant des croyances vraies se limite à celle d'un outil logique ou quasi-logique de généralisation ; la seconde consiste à conclure qu'un tel outil logique de généralisation ne joue pas de rôle explicatif-causal. Mais, poursuit Damnjanovic, si on accepte le modèle de l'explication causale développé par Frank Jackson et Philip Pettit, [54] on constate que le prédicat de vérité tel qu'il est analysé par les déflationnistes dans leur réponse à l'argument du succès, c'est-à-dire en tant que simple outil logique permettant d'exprimer des généralisations, possède toutes les caractéristiques d'une propriété causalement explicative. [55] Pour Damnjanovic, ceci n'est pas nécessairement un problème pour le déflationnisme dans la mesure où d'après le modèle de Jackson et Pettit, des notions logiques peuvent elles aussi se voir investies d'un rôle causal-explicatif. [56]

53. Voici comment Damnjanovic résume lui-même l'articulation générale de son argumentation :

> Les déflationnistes ont typiquement répondu à cet argument [N.D.T. l'argument du succès] en deux temps. Premièrement, ils tentent d'expliquer le rôle de la vérité dans ces explications sans traiter le prédicat de vérité comme autre chose qu'un outil logique bien pratique. Puis ils affirment que si le prédicat de vérité ne joue que le rôle d'un outil logique bien pratique, la vérité ne joue pas de rôle explicatif-causal.
>
> Malheureusement, la réponse standard des déflationnistes est incohérente. [...] la première étape de la réponse montre en fait la manière dont la vérité *est* une propriété causalement explicative. Toutefois, l'argument que j'examine, s'appuie sur l'explication de Jackson et Pettit des propriété causalement explicatives, et selon cette explication même les propriétés logiques peuvent se révéler être causalement explicatives. Ceci implique que le déflationniste devrait rester neutre sur la question de savoir si la vérité est une propriété causalement explicative, et se concentrer au contraire sur la thèse d'après laquelle la vérité, s'il s'agit bien d'une propriété, est simplement une propriété logique. Damnjanovic (2005, p. 54)

54. Voir Jackson et Pettit (1990a,b).

55. Plus précisément, le modèle de Jackson et Pettit a été développé dans un contexte assez éloigné du déflationnisme : celui du problème des explications causales en philosophie de l'esprit, avec pour objectif de répondre à certaines critiques d'épiphénoménalisme portant sur l'approche fonctionnaliste des états mentaux. Ce modèle distingue deux types de propriétés causales pouvant entrer dans une explication : les propriétés causalement efficaces (*causally efficacious*), qui, pourrait-on dire, sont les propriétés physiques fondamentales portant la causalité (par exemple, « avoir une structure atomique de tel ou tel type »), et les propriétés causalement pertinentes (*causally relevant*) qui « surviennent » sur les propriétés causalement efficaces et peuvent être multiréalisées par diverses propriétés causalement efficaces (par exemple, « être un bon conducteur électrique »). Damnjanovic montre que l'analyse du prédicat de vérité avancée par les déflationnistes dans leur réponse à l'argument du succès en fait précisément une propriété *causalement pertinente* au sens du modèle de Jackson et Pettit. Pour plus de détails voir Damnjanovic (2005, en particulier p. 62-66).

56. Pour une analyse et une défense détaillée du rôle causal explicatif des notions logiques, et plus

Ce dernier résultat peut paraître surprenant et même, peut-être, assez contre-intuitif. Mais nous ne chercherons pas ici à défendre la nature causalement explicative des notions logiques ni à analyser plus avant le rôle que peut jouer la vérité dans des explications causales semblables à celle déployée dans l'argument du succès. Nous avons simplement fait référence à l'article de Damnjanovic pour illustrer le fait que l'inanité explicative des notions logiques est loin d'être aussi évidente qu'il pouvait le sembler à première vue. Selon certains modèles de l'explication, ces notions peuvent avoir un rôle explicatif et même un rôle explicatif-causal (ce qui est certainement une propriété beaucoup plus forte).[57] Autrement dit, quelle que soit la manière dont ont entend interpréter les phénomènes de non-conservativité (notamment ceux impliquant des notions habituellement

particulièrement de la notion d'égalité (de grandeurs physiques entrant dans une explication causale), nous renvoyons à Damnjanovic (2005, p. 64-65), où à Jackson et Pettit (1990a, p. 207-208) pour la source originale de cet exemple.

57. Il n'en demeure pas moin que la mise en regard de Galinon (2010) et de Damnjanovic (2005) est pour le moins troublante. Les positions respectives de ces deux auteurs concernant la logicité et l'absence de pouvoir explicatif de la vérité déflationniste sont les suivantes :

D'après Galinon (2012, 2010), *la thèse centrale du déflationnisme* est que *la vérité n'est pas une notion explicative*. Ceci est censé permettre le rapprochement de la vérité déflationniste avec les notions logiques, présentées comme des cas paradigmatiques de notions non explicatives et purement expressives. La non-conservativité de ces dernières est alors évoquée pour désserrer la contrainte de conservativité portant sur la vérité déflationniste.

Pour Damnjanovic (2005), *la conception fondamentale du déflationnisme* consiste avant tout *à voir dans la vérité une sorte de notion logique*. Dès lors, toujours selon Damnjanovic (2005), le fait que les notions logiques apparaissent comme des notions dotées d'un rôle explicatif (et même d'un rôle explicatif causal) selon le modèle de l'explication causale proposé par Jackson et Pettit (1990a) devrait conduire les déflationnistes à accepter l'éventualité que la vérité puisse être une notion causalement explicative.

—L'article de Damnjanovic ne traite pas des problèmes liés à la conservativité, mais si l'on tient absolument à tracer un lien direct avec la question qui nous occupe, on peut certainement considérer sans exagérer qu'une notion explicative, dans la mesure précisément où elle nous fournit une explication, a de forte chances d'engendrer des extensions non-conservatives—

Quoi qu'il en soit, le rapprochement avec les notions logiques est mis en avant par ces deux auteurs pour soutenir des conclusions *diamétralement* opposées concernant la nature plus ou moins explicative de la vérité déflationniste.

Où l'on voit une fois encore que le « cœur » de la conception déflationniste de la vérité est décidément bien difficile à cerner, au point qu'il semble parfois être à géométrie variable, ce qui n'en facilite guère l'analyse précise et l'éventuelle critique. Mais c'est ainsi.

Qu'une chose au moins soit claire aux yeux du lecteur : en l'état actuel des arguments fournis à l'appui de l'une ou de l'autre, nous ne souscrivons ni à l'une, ni à l'autre des thèses déflationnistes suivantes :

1. la vérité est une notion logique (ou semblable aux notions logiques, ou quasi-logique ...) —sur ce point voyez le chapitre précédent.
2. la vérité n'est pas une propriété pouvant jouer un rôle explicatif au sein de nos théories —sur cet autre point voyez la discussion qui suit.

Nous n'y souscrivons pas, et ce quel que soit l'ordre de priorité qu'on veuille leur attribuer pour caractériser la doctrine centrale du déflationnisme, et quelle que soit celle à laquelle on serait le plus prompt à renoncer pour tenter de « sauver » le déflationnisme face à des résultats techniques récalcitrants.

qualifiées de logiques), on peut avoir des raisons indépendantes mais néanmoins pertinentes pour l'analyse du déflationnisme de douter du caractère non-explicatif, purement expressif, non-substantiel des notions logiques. Autrement dit encore, sans préjuger ce qu'il faut retenir de la non-conservativité, on peut avoir des raisons indépendantes de remettre en cause les intuitions sous-tendant la prémisse 2., ce qui affaiblit d'autant l'inférence ($\{1., 2.\} \vdash 3.$).

Pour finir et en guise de conclusion sur ce point, nous voudrions faire remarquer que de toute manière, même si l'on accepte l'inférence ($\{1., 2.\} \vdash 3.$), l'argument, à soi seul, ne nous semble pas suffisant pour exonérer totalement le déflationnisme d'une contrainte de conservativité. En effet, ce que montre l'argument tout au plus, c'est qu'il existe peut-être des cas « bénins » de non-conservativité. Peut-être effectivement que les notions logiques produisant des extensions non-conservatives sont des cas révélateurs prouvant que dans certaines circonstances des notions non-explicatives peuvent être non-conservatives. Mais il ne s'en suit pas que tous les cas de non-conservativité sont de cet ordre. S'il existe des cas « bénins », il existe aussi certainement des cas « graves » ou « pathologiques », c'est-à-dire des cas où la non-conservativité est bel et bien le symptome de la force explicative des notions impliquées. On peut penser à nouveau au cas des hypothèses en matière de cardinaux transfinies sur les mathématiques finitistes

La question qui se pose est alors de savoir quelle est la nature de la non-conservativité impliquant la notion de vérité, et plus spécifiquement pour le débat qui nous occupe quelle est la nature de la non-conservativité des extensions aléthiques tarskiennes telles que (PA^{+Ind}_{Tar}). S'agit-il d'un cas « bénin » similaire à celui des constantes logiques, ou au contraire d'un cas « grave » (pour le déflationnisme) ? En somme, l'argument ci-dessus nous invite simplement à la prudence dans la manière dont on doit interpréter les cas de non-conservativité. Il faudra donc examiner plus attentivement le rôle joué par le prédicat de vérité dans les extensions aléthiques non-conservatives. C'est ce que nous allons faire lorsque nous allons examiner plus en profondeur la réponse d'Hartry Field à l'argument de Shapiro et Ketland. Mais auparavant, nous voudrions dire quelques mots sur la manière dont il convient de formuler précisément la contrainte de conservativité.

4.1.5 Conservativité : peut-être, mais sur quelle théorie ?

À supposer qu'on n'écarte pas d'emblée qu'une *certaine* contrainte de conservativité pèse *peut-être* sur une axiomatisation déflationniste de la vérité, une première question

surgit rapidement ; celle de la formulation précise de cette contrainte. La conservativité est une propriété reliant deux théories formalisées. On peut donc se demander sur quelle(s) théorie(s) de départ, au juste, la vérité se doit d'être conservative ? Doit-elle l'être sur toute théorie de base, quel que soit le contenu de celle-ci ? Sur certaines seulement ? Pourquoi diable se focalise-t-on sur l'artihmétique (et même plus particulièrement sur l'arithmétique de Peano, PA), alors que c'est justement là qu'apparaissent les si problématiques phénomènes d'incomplétude gödeliens ? Pourquoi ne pas choisir une théorie logiquement plus « neutre » ? [58]

De fait, à la suite de Shapiro (1998b) et Ketland (1999), les discussions concernant la question de la conservativité se sont tout d'abord essentiellement articulées autour des phénomènes gödeliens et de la conservativité au-dessus de PA, l'arithmétique de Peano (au premier ordre). Néanmoins, si la vérité déflationniste est censée n'avoir *aucun contenu substantiel* ou *aucun pouvoir explicatif* et si l'on accepte de traduire cette caractéristique par une contrainte de conservativité, il pourrait sembler naturel d'exiger que notre théorie formelle de la vérité soit conservative au-dessus de n'importe quelle théorie de base (et non pas seulement sur PA). De façon équivalente, ceci revient à exiger que notre théorie aléthique soit conservative sur la théorie « vide », la pure logique avec identité (*i.e.* une théorie de base constituée uniquement d'axiomes et de règles de déduction logiques). [59]

Mais, dans un article paru en 2001 (Halbach, 2001b), Volker Halbach montre que pour saisir la non-substantialité de la vérité, au sens où cette notion n'aurait aucun rôle

58. Remarquons qu'à l'inverse si nous exigeons que nos théories de base soient catégoriques, par exemple en remarquant qu'une théorie qui n'est pas assez forte pour « déterminer » son modèle à isomorphisme près, est en un sens ambiguë puisqu'elle peut-être interprétée de diverses manières par diverses structures, alors la contrainte de conservativité sur ce type de théories sera trivialement satisfaite. En effet, une théorie de base catégorique est complète au sens où pour tout énoncé φ de son langage $\mathcal{L}$, $T \models \varphi$ ou $T \models \neg\varphi$. Toute extension de T sera alors (au moins sémantiquement) conservative (pour les énoncés de $\mathcal{L}$), et ce quel que soit son contenu supplémentaire par ailleurs. Le problème c'est que, du moins si on se cantonne à la logique de premier ordre, les théories catégoriques en ce sens sont bien rares et forment plutôt l'exception que la règle.

59. En effet, pour rappel, si une axiomatisation de la vérité pour les énoncés de notre langage de base $\mathcal{L}$ est conservative sur la logique pure, elle le sera sur toute théorie formulée dans $\mathcal{L}$:

Notant V les axiomes pour la vérité, supposons que V est conservative sur la logique, *i.e.* pour tout $\mathcal{L}$-énoncé φ, $V \vdash \varphi$ ssi $\emptyset \vdash \varphi$. Alors, pour toute théorie Σ formulée dans $\mathcal{L}$ et tout $\mathcal{L}$-énoncé φ, si $\Sigma \cup V \vdash \varphi$, il existe un sous-ensemble fini V_0 de V tel que $\Sigma \cup V_0 \vdash \varphi$, d'où il suit que $\Sigma \vdash v_0 \to \varphi$ (où v_0 est l'énoncé formé par la conjonction des énoncés de V_0). Comme $V \vdash v_0$ et que V est conservative sur la logique, on a $\emptyset \vdash v_0$. D'où, $\Sigma \vdash v_0$, et donc $\Sigma \vdash \varphi$. Autrement dit, $\Sigma \cup V \vdash \varphi \Rightarrow \Sigma \vdash \varphi$, pour tout φ. V est conservative sur Σ.

Cette question du caractère conservatif ou non de la vérité sur la logique pure est très importante. Elle est en particulier centrale pour l'évaluation de la logicité éventuelle d'un prédicat décitationnel. Nous avons déjà examiné cette question plus en détail dans le chapitre sur la logicité 2.

explicatif à jouer dans nos entreprises théoriques, cette exigence semble beaucoup trop forte. Il montre en effet, que la théorie minimale, composée des seules **T**-équivalences n'est déjà pas conservative sur la logique pure avec identitité. En effet, pour pouvoir simplement être formulée, cette axiomatisation de la vérité impose d'introduire toute une classe d'objets, à savoir des noms, ou des descriptions, ou des codages d'énoncés auquel le prédicat de vérité pourra s'appliquer. Ceci a pour conséquence que l'on peut alors déduire l'existence d'au moins deux objets distincts, chose qui est impossible en s'appuyant uniquement sur la logique.[60] Selon Halbach, ce résultat montre que la théorie de la vérité la plus minimale n'est déjà pas neutre ontologiquement et que donc l'exigence de conservativité (sur la logique) est tout simplement exorbitante, trop stricte et impossible à satisfaire. Cependant, Halbach poursuit son argumentation et donne un éclaircissement très utile sur la façon dont on doit considérer la contrainte de conservativité pour la vérité. Il affirme en effet que toute théorie de la vérité, dès lors qu'elle est formalisée sous une forme qui introduit un prédicat nouveau appliqué à certaines entités, présuppose en fait une théorie sous-jacente des objets auxquels le prédicat s'applique, que ces derniers soient des (noms d') énoncés, des propositions ou quoi que ce soit d'autre ... Ainsi, attribuer le prédicat « est vrai » , même s'il est « non-substantiel » ou dénote une « non propriété », sous-entend néanmoins l'existence de porteurs. Que cette ontologie sous-entendue ressorte lorsqu'on manipule notre théorie étendue n'a, au fond, rien d'étonnant. Les théories décitationnelles et néo-tarskiennes que nous avons examinées supposent toutes, a minima, l'existence d' « énoncés abstraits », ou « énoncés-types », souvent assimilés à leurs numéros de Gödel par arithmétisation de la syntaxe.

Shapiro (2003) reprend les résultats de Halbach et accepte la nécessité de prendre en compte la théorie de la syntaxe préalablement nécessaire à notre formulation de théories axiomatisant un prédicat de vérité. Il considère qu'on peut être raisonnablement neutre sur cette question en prenant comme ontologie sous-entendue de notre syntaxe, une théorie qui inclut (au moins) les suites (ou suites possibles) de symboles d'un alphabet fixé. Il rappelle en outre que les suites finies sur un alphabet fini (ou dénombrable) sont isomorphes aux entiers naturel, ce qui justifie à ses yeux l'adoption d'une théorie arithmétique élémentaire[61] comme théorie d'arrière plan fournissant les éléments syntaxiques

60. C'est cet engagement ontologique tacite qui pose particulièrement problème pour la thèse, souvent formulée de manière relâchée, selon laquelle la vérité serait une notion logique. Nous renvoyons à nouveau au chapitre où nous traitons cette question (2).

61. Par exemple PA. Mais on peut également considérer une arithmétique plus faible comme le système Q de Robinson sans induction. Ce système suffit à formaliser la syntaxe d'un langage fini ou dénombrable.

nécessaires à la formulation des théories formelles de la vérité.

Ainsi, le processus qui nous occupe se déroule en fait en trois étape :

1. on se donne une théorie objet T, énoncée dans un certain langage $\mathcal{L}$ et portant sur un certain domaine de discours (par exemple les atomes, ou les zèbres, ou les entiers naturels, ou simplement la pure logique sur un domaine non vide, *etc.*).
2. on se donne une théorie T' de la syntaxe pour les expressions de $\mathcal{L}$ (en général cette théorie sera plus ou moins équivalente à un fragment de l'arithmétique de Peano).
3. muni des instruments syntaxiques nécessaires, on ajoute à $T \cup T'$ une théorie de la vérité, disons $V(T \cup T')$ formalisant nos usages d'un prédicat de vérité en rapport avec le domaine de discours de T.

Bien sûr, si T est très « modeste », le passage de 1. à 2. peut avoir d'importantes conséquences ontologiques implicites.[62] Par contraste, si notre théorie T, prise dès la première étape, contient déjà les ressources nécessaires pour formaliser sa propre syntaxe, alors les étapes 1. et 2. se confondent ou, pour le dire autrement, le passage par 2. devient superflu. L'étape syntaxique dans le scénario ci-dessus est un problème important pour l'analyse du déflationnisme en général. Il mériterait certainement qu'on y regarde d'un peu plus près : quelle est la bonne théorie de la syntaxe ? Ou plus généralement, quelle est la bonne théorie des porteurs de vérité ? Sommes-nous seulement sûrs, d'ailleurs, que ces porteurs puissent simplement être obtenus à partir de notions purement syntaxiques ? La vérité ne porte-t-elle pas plutôt sur le contenu sémantique des énoncés[63] ? D'une manière générale la questions des porteurs de vérité est très controversée. On peut à tout le moins remarquer que les diverses théories des porteurs de vérité, que ceux-ci soient pris comme étant des propositions, ou des énoncés, types ou tokens, ont toutes des conséquences ontologiques non neutres et sont toutes soumises à l'argument de Halbach. Même le nominaliste le plus strict doit supposer l'existence d'objets, ne serait-ce que les lignes encrées sur la feuille de papier, dont l'axiomatisation sera non-conservative sur la logique pure.[64]

Sur les liens, pour le moins « intimes », entre théorie syntaxique et théorie arithmétique, nous renvoyons de nouveau à Corcoran, Frank et Maloney (1974).

62. C'est ce que met bien en lumière l'argument de Halbach, 2001b.

63. Nous avons déjà évoqué et quelque peu discuté ce problème dans le chapitre précédent (2.4.2.4). Comme cette question n'est pas directement centrale pour la discussion de l'argument de Keltand et Shapiro, nous ne nous y attardons pas ici.

64. Et c'est dans le contexte de la discussion de l'éventuelle logicité d'un prédicat déflationniste de

Ceci étant, pour ce qui est de mesurer la substantialité de la notion de vérité, au sens où ce concept pourrait avoir un rôle explicatif à jouer dans nos discours théoriques sur le monde, c'est, nous semble-t-il, le passage de 2. à 3. qui est primordial. Si la vérité s'avère conservative sur une théorie de la syntaxe du langage considéré, alors il y a bien un sens à affirmer que ce prédicat n'augmente en rien notre pouvoir explicatif et nos capacités à rendre compte des phénomènes au travers de nos théories. Ainsi, nous souscrivons à la conclusion de Halbach, 2001b et à la nécessité de préciser et reformuler quelque peu le critère de conservativité en jeu :

> « ... si la vérité est non substantielle, alors elle ne devrait rien impliquer au-delà des présupppositions faites lors de sa formulation » [65]

Si la vérité doit être conservative au-dessus d'une « théorie raisonnable de la syntaxe » (indispensable à la simple possibilité de formuler une théorie de la vérité [66]), alors la conservativité sur une théorie objet contenant déjà sa propre syntaxe est bien la question fondamentale. À ce titre, considérer directement une théorie de base contenant un minimum d'arithmétique apparaît simplement commode et philosophiquement justifié.

Pour finir sur ce point, nous voudrions faire une dernière remarque. Si par une « théorie raisonnable de la syntaxe » on entend une théorie capable d'établir, c'est-à-dire prouver, un certains nombres de faits concernant la syntaxe du langage de la théorie de base, alors sans doute faut-il s'attendre à ce que cette « théorie raisonnable de la syntaxe » vérifie certaines propriétés de représentabilité —du type de celle exposées dans le chapitre précédent. [67] Dès lors, elle sera soumise aux théorèmes d'incomplétude de Gödel. C'est pourquoi, les phénomènes gödeliens sont vraiment le lieu central et « naturel » pour la discussion de l'argument de la conservativité. Le choix de ce type de théories n'a rien d'hasardeux ou de particulièrement défavorable au déflationnisme. Pour le dire autrement, il ne s'agit pas de sélectionner machiavéliquement et à dessein une théorie de base incomplète au sens de Gödel parce qu'on pressent que ce type de théorie pourra poser problème aux déflationnistes. Il s'agit plutôt de choisir une « théorie raisonnable de la syntaxe », de toute manière nécessaire pour pouvoir formuler correctement une axiomatisation de la vérité, puis de constater que cette théorie est soumise aux phénomènes

vérité que cette non-conservativité-*là*, *i.e.* la non-conservativité sur la logique pure, prend toute son acuité. *Cf.* 2.

65. Halbach, 2001b, p. 182.

66. Du moins, rappelons le, si on entend par là une théorie axiomatisant le *prédicat* « vrai », prédicat qui devra donc trouver des objets auxquels s'appliquer.

67. 3.3.1.2.

d'incomplétude et d'en tirer les conséquences pour un examen des thèses déflationnistes. Ainsi, à la fin du paragraphe précédent au lieu d'évoquer une théorie de base contenant un minimum d'arithmétique, nous aurions pu tout aussi bien écrire : ...
« À ce titre, considérer directement une théorie de base capable de réprésenter sa propre syntaxe et soumise, de ce fait, aux phénomènes d'incomplétude gödeliens apparaît simplement commode et philosophiquement justifié ».
Ce qui est important pour l'analyse de l'argument de la conservativité, ce n'est pas que nous parlions d'arithmétique, mais bien que nous prenions comme point de départ une théorie capable de représenter sa propre syntaxe et qui sera, de ce fait, soumise aux phénomènes d'incomplétude.

4.2 Schéma d'induction et extensions aléthiques

Une fois précisé la manière dont la contrainte de conservativité doit être formulée, et plus particulièrement le type de théorie de base sur laquelle elle doit éventuellement s'exercer, il reste à évaluer plus précisément la portée des résultats de non conservativité exhibés par Shapiro et Ketland pour ce qui concerne l'évaluation du déflationnisme. Puisque les déflationnistes revendiquent l'indispensabilité expressive du prédicat « vrai » mais insistent sur son caractère non explicatif et non substantiel, il nous faut notamment sonder plus en profondeur le rôle que joue la vérité dans les preuves sémantiques accompagnant les phénomènes d'incomplétude. L'un des points cruciaux pour les arguments qui vont suivre concerne le rôle du schéma d'induction dans l'apparition des phénomènes de non conservativité. Comme nous l'avons rappelé dans le chapitre précédent, lorsqu'on part d'une théorie de base arithmétique contenant un schéma d'axiomes d'induction, comme par exemple l'arithmétique de Peano (PA) au premier ordre, et qu'on étend cette théorie au moyen d'axiomes pour la vérité, les résultats de conservativité vont varier selon l'extension aléthique considérée. Une extension aléthique limitée aux seuls axiomes de la décitation, $\{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{PA}\}$, est trop faible pour nous permettre de prouver de nouveaux théorèmes du langage de base $\mathcal{L}_{PA}$, et ce, même si on s'autorise à étendre le schéma d'induction aux énoncés du langage étendu $\mathcal{L}' = \mathcal{L}_{PA} \cup \{Vr\}$. Autrement dit, PA_d et PA_d^{+Ind} sont toutes deux conservatives sur PA. À l'inverse, une extension aléthique contenant les axiomes d'une théorie compositionnelle à la Tarski va, sous certaines conditions, engendrer une extension non conservative en permettant de dériver un principe de réflexion pour l'arithmétique de Peano et donc de mener une preuve

sémantique de consistance. Mais, et c'est là le point capital, l'extension tarskienne ne sera non conservative[68] que si l'on autorise l'emploi de l'induction portant sur des énoncés qui contiennent le prédicat « vrai », c'est-à-dire portant sur de énoncés qui contiennent du vocabulaire du langage étendu[69] Autrement dit, l'extension aléthique tarskienne ne devient non conservative que si l'on traite le schéma d'induction comme une *règle* et non comme une *liste*[70] : PA_{Tar} reste conservative sur PA, et ce n'est qu'en passant à PA^{+Ind}_{Tar} que l'on obtient une théorie de la vérité réflexive. [71] Ce point technique subtil est au coeur de certaines défenses du déflationnisme face au dilemme proposé par Shapiro et Ketland. Étant donné l'importance de l'induction étendue au vocabulaire sémantique dans l'apparition de la non-conservativité, on peut à première vue deux imaginer deux lignes de défenses pour le déflationniste : la première pourrait consister à refuser de considérer l'induction comme *règle* dans le cas d'une extension par des principes aléthiques et à se cantonner à un schéma compris comme une *liste* d'axiomes ne contenant que des énoncés du langage de la théorie de base. Cette première thèse n'a cependant guère été ouvertement défendue par les déflationnistes[72] La seconde consiste à accepter l'élargissement de l'induction mais à en donner une interprétation qui « préserve » la vérité déflationniste. Cette seconde ligne de défense a en revanche été ouvertement mise en avant par Field (1999) en réponse aux arguments de Ketland (1999) et Shapiro (1998b) et à engendrée une riche discussion. C'est donc sur cette seconde ligne de défense que nous allons nous concentrer ici.

4.2.1 L'induction et les axiomes « essentiels » de la vérité

Examinons à présent la seconde stratégie de défense déflationniste articulée autour du rôle particulier que joue l'extension du schéma d'induction dans la formalisation des arguments sémantiques et l'émergence des phénomènes de non conservativité. Cette fois, la manœuvre ne consiste pas à refuser l'emploi de l'induction portant sur le vocabulaire sémantique pour bloquer l'apparition de la non conservativité. Il s'agit plutôt

68. Au sens syntaxique, et donc au sens sémantique, dans le cadre d'une logique du premier ordre
69. Lequel vocabulaire n'est bien sûr pas réductible au moyen d'une définition au langage de base, d'après le célèbre théorème d'indéfinissabilité de la vérité de Tarski.
70. Selon la terminologie empruntée à Burgess et Rosen (1997).
71. Au sens de Ketland (1999) et Shapiro (1998b).
72. Cette position a néanmoins été attribuée à Tennant par Shapiro (2003). Voyez la note 6, page 264. Azzouni (1999) développe quant à lui une position d'après laquelle le déflationniste devrait se cantonner à une théorie conservative sur PA et donc refuser d'étendre l'induction. Pour plus de détails sur cette ligne de défense déflationniste et ses variantes, et pour une réponse , voyez Shapiro (2003).

ici d'examiner plus attentivement la structure des arguments sémantiques menant à des démonstrations de consistance dans l'intention de montrer qu'en dépit de la non conservativité le rôle qu'y joue la notion de vérité est compatible avec la conception que s'en font les déflationnistes.

Nous avons déjà évoqué les fait que certains auteurs déflationnistes entendaient distinguer rôle expressif d'une part, et pouvoir explicatif ou substantialité d'autre part. Et nous avons aussi rappelé qu'ils ne reconnaissaient à la vérité que la première de ces deux caractéristiques. Ces auteurs insistent en outre sur l'indispensabilité du prédicat de vérité en tant qu'outil de généralisation permettant d'accroître les capacités expressives de notre langage. Une fois établie cette distinction et rappelée cette indispensabilité, il semble envisageable que la notion de vérité puisse se présenter dans une théorie simplement en raison de la fonction expressive qu'elle remplit et sans pour autant être investie d'un quelconque rôle explicatif ou d'une quelconque substantialité. Ce pourrait notamment être le cas, peut-être, dans une théorie aléthique étendant une théorie arithmétique de base, même si cette extension s'avère être non conservative. Nous avons également déjà dit qu'à nos yeux, ce type d'argumentation nécessiterait pour être pleinement menée à bien de clarifier la distinction entre rôle expressif et pouvoir explicatif ou substantialité, et surtout de donner un critère précis permettant de tracer la frontière entre les notions purement expressives et les notions substantielles ou explicatives.[73] À notre connaissance, une telle distinction générale appuyée sur un critère précis de cet ordre n'a jamais été fournie par les déflationnistes.

Pour autant, la réponse proposée par Field (Field, 1999) à l'argument de Shapiro et Ketland peut être vue comme une tentative de défense du déflationnisme s'inscrivant dans cette ligne et appliquée au cas particulier des extensions tarskiennes de l'arithmétique de Peano. En somme, l'objectif de Field est de montrer que la non conservativité de l'extension PA^{+Ind}_{Tar} où l'induction est étendue aux énoncés contenant le prédicat « vrai » est compatible avec les thèses déflationnistes concernant la nature de la vérité, en défendant l'idée que, contrairement à ce que pourrait laisser penser la non conservativité, le prédicat de vérité ne joue pas de rôle explicatif dans la preuve sémantique de consistance.

73. ou peut-être entre les fonctions expressives d'une notion et ses fonctions explicatives ou ses emplois substantiels, lesquelles fonctions pourraient parfois se combiner et s'exercer simultanément quand, à d'autres occasions, elles seraient tout à fait séparées et employées de manière indépendante, avec pour cas limite les situations où une notion ne remplirait qu'une fonction « purement expressive ».

4.2.1.1 La position de Field (1999)

Rappelons les principaux aspects de la position de Field telle qu'elle se dégage de son article de 1999. Concernant la contrainte de conservativité, nous avons déjà eu l'occasion de dire que Field semble ici tout proche de l'accepter :

> Étant donné que la vérité peut être ajoutée d'une manière qui engendre une extension conservative (même en logique du premier ordre), il n'est pas nécessaire d'être en désaccord avec Shapiro lorsqu'il dit : « la conservativité est essentielle au déflationnisme » (Field, 1999, p. 536, 1er §)

S'il paraît donc accepter la condition de conservativité, Field affirme néanmoins qu'un prédicat de vérité non-compositionnel, au sens où il n'obéirait pas aux clauses tarskiennes, serait logiquement insatisfaisant et inutile. Il ne semble donc pas souscrire à une théorie purement décitationnelle du type de PA_d avec ou sans induction étendue puisqu'il déclare :

> [...] il est plus intéressant d'ajouter la vérité d'une manière qui comprenne les lois générales [N.D.T : ce que Field désigne ici par lois générales, ce sont des axiomes compositionnels à la Tarski, semblables à ceux que nous avons notés Tar [74]], puisque je pense qu'il est clair qu'en l'absence de telles lois générales le prédicat de vérité ne pourrait pas remplir sa principale fonction [N.D.T. : à savoir permettre de formuler des généralisations] (Field, 1999, p. 535) [75]

Ainsi, *a minima*, Field semble favoriser une axiomatisation néo-tarskienne de la vérité du type PA_{Tar}. Mais il va même plus loin, puisque lorsqu'il analyse les arguments de Shapiro (1998b), il écrit que

> les axiomes d'induction [N.D.T. : étendus aux formules contenant le prédicat de vérité.] sont nécessaires si nous voulons dériver arithmétiquement des faits importants concernant la vérité. (Field, 1999, p. 538),

74. Voyez le chapitre précédent pour une formulation détaillée : 3.3.2.3.

75. En outre, dans une note Field reprend à son compte la critique de Gupta (1993) contre la formulation déflationniste limitée aux seules instances des **T**-équivalences :

> Je concède qu'un déflationniste, Paul Horwich —*Truth*— a formulé sa théorie de la vérité au moyen de principes qui ne sont pas assez forts pour donner à « vrai » le rôle expressif que nous voulons qu'il ait. Il a été correctement critiqué pour cela par [Gupta, 1993]. (Field, 1999, p. 534, note 4).

ou encore que

> [les axiomes d'induction sont importants] parce qu'il est certain que nous voulons être capables de prouver inductivement que tous les théorèmes de T sont vrais sur la base de la vérité des axiomes et de la préservation de la vérité par les règles d'inférences. (Field, 1999, p. 538)

Ce dernier passage, insisitant sur la nécessité de pouvoir dériver un schéma de réflexion pour T, ressemble fort à une acceptation en bonne et due forme de la contrainte de réflexivité.[76] Quoi qu'il en soit, Field semble résolument disposé à étendre le schéma d'induction aux énoncés contenant le prédicat de vérité.[77] Au total, Field souscrit donc visiblement à une théorie compositionnelle de la vérité proche de PA^{+ind}_{Tar}. Mais, bien sûr, PA^{+ind}_{Tar} *n*'est *pas* conservative ! Comment Field résout-il l'antinomie apparente de sa position ?

La justification de Field repose sur une distinction assez subtile. Selon lui, au cours de l'opération qui, ayant pour point de départ l'arithmétique du premier ordre comme théorie de base, nous conduit à accepter PA^{+ind}_{Tar}, il convient de distinguer ce qu'il appelle les « axiomes essentiels de la vérité » et les axiomes qui concernent en fait les entiers naturels. Les axiomes « essentiels » de la vérité sont ceux dont le contenu, tout comme la formulation et la justification, ne dépend que de la nature propre de la vérité. Pour Field, la contrainte de conservativité traduisant l'absence de substantiatlité et de pouvoir explicatif allégués par les déflationnistes, ne doit justement s'exercer que sur ces axiomes essentiels : rien d'étonnant après tout à ce que si l'on considère des axiomes dont le contenu ne relève pas uniquement et totalement de la vérité, on puisse voir apparaître des phénomènes de non conservativité.

Parmi ces axiomes « essentiels », il faut compter les clauses récursives composition-

76. Dans ce passage de son article, Field cite librement les arguments de Shapiro. Il n'est donc pas tout à fait clair qu'il accepte lui-même la contrainte de réflexivité. Pour autant, cette interprétation cadre parfaitement avec le reste de son argumentation (en particulier avec le fait que Field adopte ici une théorie formelle de la vérité semblable à PA^{+ind}_{Tar}). En outre, Field ne signale à aucun moment son désaccord avec *cet* argument précis de Shapiro (*i.e.* l'argument selon lequel nous voulons être capables de prouver inductivement quelque chose comme (Ref) pour notre théorie de base), alors qu'il signale clairement son désaccord avec d'autres arguments. On peut donc raisonnablement considérer que Field accepte la contrainte de réflexivité.

77. D'ailleurs, Field justifie l'extension du schéma d'induction arithmétique, pris comme *règle*, au vocabulaire sémantique, en s'appuyant sur une argumentation proche de celle de Shapiro : à la fin de son article Field reconnaît que lorsqu'on s'engage envers le schéma d'induction arithmétique, on entend s'engager non seulement envers les instances du langage environnant mais également envers toutes les instances de toute expansion légitime de ce langage, notamment celle contenant le prédicat de vérité (voir Field (1999, p. 539)).

nelles, ce que Field appelle les règles générales, gouvernant l'emploi du prédicat « vrai » et qui correspondent *grosso modo* à la collection d'axiomes exprimés dans $\mathcal{L}_{PA} \cup \{Vr\}$ que nous avons notée Tar. En revanche, nous dit Field, nous n'avons aucune raison de considérer les instances du schéma d'induction comportant le prédicat de vérité comme faisant partie de ces « axiomes essentiels ». Field reconnaît volontiers l'importance de l'induction étendue dans les arguments sémantiques :

(a) cette induction étendue est nécessaire si nous voulons pouvoir dériver arithmétiquement certains faits importants concernant la notion de vérité.

Mais de là, il ne suit pas que

(b) les nouvelles instances du schéma d'induction étendu à la vérité dépendent *uniquement* de la nature de la vérité. [78]

Or, pour Field, c'est bien quelque chose comme (b) qui serait nécessaire pour pouvoir classer les instances de l'induction portant sur le prédicat « vrai » parmi les axiomes essentiels de la vérité. Et (b) semble évidemment faux : ce qui justifie l'extension de l'induction à un nouveau vocabulaire, sémantique ou autre, c'est, au moins pour partie, la nature et la structure des nombres entiers. Il n'y a là rien qui fasse intervenir la nature propre de la vérité ; nous pourrions tout aussi bien étendre le schéma d'induction à tout autre vocabulaire nouveau désignant une propriété portant sur les entiers. Dès lors, toujours selon Field, dans l'élargissement de l'induction, *i.e.* le passage de PA_{Tar} à PA^{+ind}_{Tar}, la vérité ne joue aucun rôle explicatif mais apparaît bien comme un simple « auxiliaire expressif » qui nous permet de formaliser une théorie arithmétique plus forte. Dans un passage crucial de son article, Field écrit par exemple :

> La façon dont nous "apprenons de nouvelles choses à propos des nombres naturels en invoquant la vérité" est qu'en disposant de cette notion nous pouvons formuler rigoureusement une théorie arithmétique plus forte que celle que nous pouvions rigoureusement formuler auparavant. Il n'y a rien ici de véritablement spécial à propos de la vérité : en employant n'importe quelle autre notion non exprimable dans le langage d'origine, nous pouvons obtenir de nouvelles instances du schéma d'induction et bien souvent ces dernières déboucheront sur des extensions non conservatives. (Field, 1999, p. 536)

Ainsi, la compositionnalité de prédicat « Vr » est déjà parfaitement saisie par les clauses tarskiennes $\{Tar\}$, qui constituent en réalité l'ensemble des « axiomes essentiels de la

78. Voir Field (1999, p. 538).

vérité ». Il est alors agréable pour le déflationniste de constater que PA_{Tar} est bien une extension conservative de PA. Ce n'est que lorsque nous passons de PA_{Tar} à PA^{+ind}_{Tar}, en élargissant notre schéma d'axiomes d'induction au vocabulaire aléthique de $\mathcal{L}' \supseteq \mathcal{L}_{PA} \cup \{Vr\}$, que nous brisons la contrainte de conservativité. Mais l'élargissement du schéma d'axiome d'induction aux énoncés contenant « Vr » se fait non pas au nom de la nature ou de l'essence de la vérité mais tout simplement en raison de faits arithmétiques concernant les entiers. En conséquence, la non-conservativité de PA^{+ind}_{Tar} n'est pas due à une nature « substantielle » de la vérité ou à un quelconque pouvoir explicatif de cette notion, et elle ne saurait remettre en cause le caractère purement expressif du prédicat de vérité.

Grâce à sa brillante combinaison, Field semble être en mesure de réconcilier les deux branches du dilemme posé par Ketland et Shapiro : la non-substantialité de la vérité et son absence de rôle explicatif se traduisent bien par un résultat de conservativité. Simplement, ce résultat concerne les « axiomes essentiels de la vérité », qui sont bien ceux sur lesquels la contrainte de conservativité doit s'exercer. La contrainte ainsi formulée et précisée est bien satisfaite puisque PA_{Tar} étend conservativement PA. Mais, le prédicat de vérité augmente également les capacités expressives de notre langage et nous permet de formuler de nouveaux axiomes d'induction. Dans cet élargissement, le prédicat de vérité ne joue pas de rôle explicatif, mais apparaît simplement comme un outil expressif de généralisation. Ces nouveaux axiomes qui débouchent sur PA^{+ind}_{Tar} ne sont pas des axiomes essentiels de la vérité et ne sont donc pas soumis à la contrainte de conservativité. Bien au contraire, ils sont importants puisqu'ils nous permettent de formuler une théorie arithmétique plus forte que celle que nous pouvions formuler en l'absence de cette nouvelle ressource expressive qu'est le prédicat de vérité. Cette nouvelle théorie arithmétique PA^{+ind}_{Tar} permettra de dériver arithmétiquement certains faits importants concernant la vérité, et satisfera la contrainte de réflexivité. Le déflationniste peut donc à la fois adopter PA^{+ind}_{Tar} et se doter des moyens de démontrer G ou $Con(PA)$, tout en revendiquant le caractère conservatif, et donc non substantiel ou non explicatif, de la vérité *stricto sensu*.

L'argumentation de Field est d'une grande ingéniosité. Elle peut paraître plausible à première vue et séduisante aux yeux d'un partisan du déflationnisme. Néanmoins, nous pensons qu'elle est en réalité vouée à l'échec et qu'elle ne résiste pas à un examen plus approfondi. Tel que nous le comprenons, le raisonnement de Field vise à établir

que les arguments formalisés dans l'extension aléthique PA^{+ind}_{Tar} sont compatibles avec la conception déflationniste de la vérité en montrant que, dans ces arguments, rien ne repose sur une notion « substantielle » de vérité mais qu'au contraire le prédicat de vérité ne joue ici qu'un rôle purement expressif. À l'appui de cette thèse, Field invoque une analyse plus fine de la structure des arguments sémantiques et qui ne se borne pas au simple constat de la non-conservativité de PA^{+ind}_{Tar} sur PA : parmi les (nouveaux) axiomes employés pour établir les schéma de réflexion puis les énoncés G ou $Con(PA)$, Field distingue les axiomes essentiels de la vérité qui sont conservatifs, tandis que dans l'extension de l'induction, nécessaire pour obtenir une preuve de G ou de $Con(PA)$, le prédicat de vérité apparaît seulement comme un auxiliaire expressif. Malheureusement cette brillante reconstruction ne résiste guère à une analyse plus poussée de la structure des arguments sémantiques.

4.2.2 Contre-arguments

4.2.2.1 Contre-argument n°1 (Heck Jr (2018) Horsten (2011), Ketland (2010) Strollo (2013))

Un premier contre-argument [79] consiste simplement à remarquer que, contrairement à ce que pourrait laisser penser l'exposé donné par Field dans son article, les axiomes compositionnels tarskiens $\{Tar\}$ jouent bel et bien un rôle crucial et indispensable dans les arguments sémantiques avancés par Shapiro et Ketland. S'il est exact qu'ils ne *suffisent* pas à eux seuls à obtenir le principe de réflexion (puisque $PA_{Tar}= PA \cup \{Tar\}$ est conservative sur PA), *a contrario* il faut toutefois souligner qu'ils sont rigoureusement *nécessaires* pour y parvenir.

D'une certaine façon, c'est justement ce qu'atteste la conservativité de PA_d sur PA. Plus précisément, lorsqu'on part de PA et qu'on l'étend au moyen des seules **T**-équivalences pour obtenir PA^{+ind}_{d}, on obtient déjà ce faisant un nouvelle ressource non exprimable dans le langage d'origine, puisqu'il est bien connu que le prédicat « Vr » ainsi caractérisé par les axiomes $\{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{PA}\}$ n'est pas définissable dans $\mathcal{L}_{PA}$. Une fois munis des **T**-équivalences , nous sommes donc déjà en possession d'un terme, à savoir « Vr », désignant un sous-ensemble d'entiers naturels non exprimable (au

79. Ce type contre-argument que nous présentons ici n'est pas nouveau. On le trouve, par exemple, sous diverses variantes dans Horsten (2011, chapitre 7) , dans Heck Jr (2018, p. 2-3) ou déjà dans Heck Jr (manuscript non publié), ainsi que dans Strollo (2013, p. 538-539) et dans Ketland (2010, p. 11 note 17).

sens de non-définissable) dans $\mathcal{L}_{PA}$.[80] Si l'on suit l'analyse de Field (1999, p. 536),[81] nous disposons donc déjà dans PA_d^{+ind} d'une notion nouvelle non exprimable dans le langage d'origine au moyen de laquelle nous pouvons obtenir de nouvelles instances du schéma d'induction et formuler ainsi une théorie arithmétique plus forte que celle que nous pouvions rigoureusement formuler auparavant. Pourtant lorsqu'on fait cela, c'est-à-dire lorsqu'on passe de

$$PA_d^{+ind} = PA \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{PA}\} + SI(\mathcal{L}_{PA})$$ [82]

à

$$PA_d = PA \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{PA}\} + SI(\mathcal{L}_{Vr})$$

80. Au demeurant, pour ce qui est de caractériser l'extension de « Vr » *sur les entiers naturels*, le choix des **T**-équivalences ou de $\{Tar\}$ pour axiomatiser le prédicat de vérité importe peu —c'est plutôt sur les entiers non standard dans les modèles non standard que la distinction s'avère cruciale. Précisons ce que nous voulons dire par là : si l'on considère $\mathcal{L}_{PA}$ comme un langage interprété au sens où on considère que les énoncés de ce langage sont accompagnés de leurs significations usuelles et nous parlent des « vrais » entiers, ce que l'on explique parfois en disant que $\mathcal{L}_{PA}$ est implicitement muni de son modèle standard $\mathfrak{N} := \langle \mathbb{N}, 0^{\mathbb{N}}, +^{\mathbb{N}}, .^{\mathbb{N}}, S^{\mathbb{N}} \rangle$, c'est-à-dire l'ensemble des entiers naturels munis du zéro, de l'addition, de la multiplication et de la fonction successeur habituels, alors prendre comme axiomatisation pour « Vr » l'ensemble des **T**-équivalences (pour obtenir PA_d^{+ind}) ou les clauses compositionnelles tarskiennes $\{Tar\}$ (pour obtenir PA_{Tar}) ne change pas grand choses concernant la caractérisation de l'extension de « Vr » : *sur le modèle standard* $\mathfrak{N}$ les deux axiomatisations nous donnent la même *unique* extension possible pour « Vr », à savoir l'ensemble des (codes des) énoncés φ de $\mathcal{L}_{PA}$ vrais dans $\mathfrak{N}$ (*i.e.* , tels que $\mathfrak{N} \models \varphi$, pour reprendre la notation habituelle de la théorie des modèles). En ce sens, on peut dire que les **T**-équivalences suffisent à « fixer », ou « caractériser », ou « déterminer » l'extension du prédicat de vérité...sur le modèle standard. Néanmoins, cela ne signifie pas que les **T**-équivalences , (pas plus que les axiomes tarskiens $\{Tar\}$), suffisent pour fixer l'extension de « Vr » sur tous les modèles possibles de PA, c'est-à-dire sur toutes les structures d'interprétation de $\mathcal{L}_{PA}$ qui rendent vrais les axiomes de Peano. Bien au contraire, il existe des modèles non standard $\mathfrak{M}$ de PA sur lesquels on peut trouver plusieurs sous-ensembles $Vr^{\mathfrak{M}}$ interprétant « Vr » de manière à satisfaire toutes les **T**-équivalences (ou à satisfaire les clauses $\{Tar\}$). Si tel n'était pas le cas, alors l'axiomatisation de « Vr » par les **T**-équivalences en donnerait ce qu'on appelle une définition implicite (au sens de Beth, c'est-à-dire une axiomatisation fixant pour *chaque* modèle possible une unique extension possible) ; et d'après le théorème de Beth, il s'en suivrait que « Vr » est explicitement définissable dans $\mathcal{L}_{PA}$, ce qui est notoirement connu pour être faux depuis le théorème d'indéfinissabilité de Tarski. En résumé, les **T**-équivalences et *a fortiori* les axiomes compositionnels $\{Tar\}$ suffisent à « fixer » l'extension de « Vr » sur le modèle standard, mais ni l'une ni l'autre de ces axiomatisations ne fixent l'extension de « Vr » sur tous les modèles de PA. En outre, il faut remarquer que sur les modèles non standard, où l'extension possible de « Vr » n'est pas nécessairement unique, l'axiomatisation $\{Tar\}$ fait peser des contraintes beaucoup plus fortes sur l'extension possible de « Vr » que ne le fait la collection des **T**-équivalences . C'est en partie ce qui est à l'origine de la force preuve-théorique plus forte de PA_{Tar}^{+ind} par rapport à PA_d (voyez sur ce point les discussions suivantes sur les classes de satisfactions).

Voir aussi Bays (2009) et Ketland (2009) sur la notion de « définition implicite » du prédicat « Vr », en un sens différent, non technique, de celui habituellement employé en théorie des modèles et notamment dans le théorème de Beth.

81. Voyez la citation tirée de son article page 298.

82. Nous utiliserons parfois cette notation $SI(\mathcal{L})$ pour préciser à quel langage le schéma d'induction est censé s'appliquer.

la « théorie arithmétique plus forte que celle que nous pouvions rigoureusement formuler auparavant » n'est cependant pas suffisamment forte pour pouvoir mener à bien les arguments sémantiques permettant d'établir $(Ref), G$, ou $Con(PA)$. Bien au contraire, les « axiomes essentiels » $\{Tar\}$ font ici défaut ; ce n'est qu'en les ajoutant, *i.e.* en passant de

$$PA_d = PA \cup \ \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{PA}\} + SI(\mathcal{L}_{Vr})$$

à

$$PA_d{+}\{Tar\} = PA \cup \ \{Tar\} + SI(\mathcal{L}_{Vr}) = PA^{+ind}_{Tar}$$ [83]

qu'on obtient une extension aléthique réflexive.

Pour le dire autrement, la reconstruction de l'argument sémantique proposée par Field exige qu'on procède par étapes et *dans l'ordre suivant* : partant d'une théorie arithmétique du premier ordre du type PA exprimée dans $\mathcal{L}_{PA}$, on introduit d'abord un nouveau symbole de prédicat unaire, à savoir « Vr » (on enrichit le langage $\mathcal{L}_{PA}$ pour obtenir $\mathcal{L}' = \mathcal{L}_{PA} \cup \{Vr\}$), ainsi que les axiomes essentiels censés en gouverner l'usage, identifiés ici aux clauses compositionnelles $\{Tar\}$; on obtient alors l'extension conservative PA_{Tar}. Puis, dans un second temps, nous tirons profit de l'augmentation de nos ressources expressives due à l'enrichissement du langage au moyen d'un prédicat inexprimable dans le langage d'origine pour formuler une théorie arithmétique plus forte et non conservative sur PA_{Tar}, à savoir PA^{+ind}_{Tar}. Dans cette seconde étape le prédicat de vérité ne joue pas de rôle explicatif, il apparaît simplement comme « auxiliaire expressif » permettant de renforcer notre schéma d'induction pour formuler une théorie arithmétique plus forte. Le processus peut donc se schématiser comme ceci :

$$PA \overset{1}{\Longrightarrow} \ PA_{Tar} \overset{2}{\Longrightarrow} PA^{+ind}_{Tar} \qquad (A)$$

Mais on pourrait tout aussi bien proposer la reconstruction suivante : partant de PA on enrichit $\mathcal{L}_{PA}$ au moyen d'un nouveau prédicat « Vr » et on étend PA en lui ajoutant la collections des **T**-équivalences $\{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{PA}\}$, laquelle collection suffit à nous donner l'extension sur $\mathbb{N}$ d'un nouveau sous-ensemble d'entiers naturels qui n'était

83. Rappelons que comme PA_{Tar} prouve toutes les **T**-équivalences , on a bien $PA_d \subseteq PA^{+ind}_{Tar}$, et donc $PA_d{+}\{Tar\} = PA^{+ind}_{Tar}$.

pas définissable dans le langage d'origine ; munis de cette nouvelle ressource expressive (*i.e.* un prédicat dont l'extension n'est pas « exprimable dans le langage d'origine »), nous formalisons une théorie arithmétique plus forte en étendant l'induction. Nous obtenons $PA_d = PA \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{PA}\} \cup SI(\mathcal{L}_{Vr})$. Cette théorie arithmétique plus forte dans laquelle, si l'on accepte une analyse à la Field, le prédicat « Vr » n'a pas de rôle explicatif, n'est cependant pas suffisante. Pour pouvoir dériver le schéma de réflexion et donner une preuve sémantique de G ou de $Con(PA)$, il nous faut en effet encore nous munir des axiomes compositionnels. On obtient donc la séquence suivante :

$$PA \overset{1}{\Longrightarrow} PA_d \overset{2}{\Longrightarrow} PA^{+ind}_{Tar} \qquad (B)$$

Selon qu'on introduit les axiomes compositionnels tarskiens $\{Tar\}$ avant ou après avoir étendu l'induction au vocabulaire sémantique, on peut leur attribuer une propriété de conservativité (sur PA en obtenant PA_{Tar} avant d'étendre l'induction pour obtenir PA^{+ind}_{Tar}) ou au contraire de non conservativité (sur PA_d, après avoir préalablement élargi l'induction). Mais quoi qu'il en soit, comme l'illustre la séqence (B) ci-dessus, ces axiomes compositionnels s'avèrent bel et bien indispensables pour dériver le principe de réflexion (Ref) et mener à son terme la preuve sémantique de G (ou de $Con(PA)$). Field a certes parfaitement raison d'affirmer que l'élargissement du schéma d'induction à des nouvelles notions non exprimables dans le langage d'origine peut *parfois* déboucher sur une théorie arithmétique plus forte non conservative et que ce phénomène n'est pas propre à la vérité. Mais, comme le souligne Heck (Heck Jr, 2018, p 2-3), ce qui est justement frappant dans le cas de la vérité, c'est que le renforcement de l'induction au moyen de la nouvelle ressource expressive « Vr » *ne suffit pas* à obtenir une théorie non conservative. Pour pouvoir dériver (Ref), nous avons besoin de prendre en compte (et d'employer sous la forme d'axiomes) une information supplémentaire sur la nature du prédicat de vérité. Cette information supplémentaire porte sur la manière dont est structurée l'extension du prédicat de vérité : pour obtenir (Ref), nous devons par exemple pouvoir nous appuyer sur le fait que pour toute paire d'entiers, si le premier entier est $\ulcorner\varphi\urcorner$, *i.e.* le code de la formule φ, et que le second est $\ulcorner\psi\urcorner$, le code de ψ et si ces deux entiers sont dans l'extension de « Vr », alors il s'en suit qu'un troisième entier [84], à savoir $\ulcorner\varphi \wedge \psi\urcorner$ le code

84. qu'on peut d'ailleurs obtenir à partir des deux précédents par l'intermédiaire d'une opération récursive.

de $\varphi \wedge \psi$ sera dans l'extension de « Vr ». De même, nous devons pouvoir nous appuyer sur le fait que si un entier $\ulcorner\varphi\urcorner$ est dans l'extension de « Vr », alors pour toute formule ψ, l'entier $\ulcorner\varphi \vee \psi\urcorner$ sera dans « Vr », *etc.* , *etc.* Autrement dit, nous devons pouvoir nous appuyer sur la nature compositionnelle du prédicat de vérité, et c'est précisément cette information qui est couchée dans les axiomes tarskiens.

À la séquence (A) mise en avant par Field, on peut donc opposer la séquence (B) qui met en lumière l'importance fondamentale des axiomes $\{Tar\}$ dans les arguments sémantiques : s'il s'agissait simplement de renforcer l'induction en l'employant sur un nouveau sous-ensemble d'entiers naturels[85], alors la non conservativité devrait apparaître dès l'élargissement de PA_d^{+ind} à PA_d. Si à l'inverse les axiomes « essentiels » de la vérité $\{Tar\}$ sont indispensables (conjointement à l'induction élargie) pour mener à bien la preuve sémantique, alors il semble difficile de les exonérer de tout rôle explicatif et de toute responsabilité dans l'apparition de la non conservativité accompagnant l'extension aléthique PA_{Tar}^{+ind}. Voilà, pour l'essentiel, en quoi consiste le contre-argument numéro 1.

On peut néanmoins en poursuivre le développement en inspectant « ligne à ligne » la dérivation de (Ref) dans PA_{Tar}^{+ind}.[86] Lorsqu'on se livre à un tel examen, on ne peut qu'être frappé de l'importance du rôle joué par les axiomes compositionnels. La preuve de l'énoncé (Ref) procède en effet par induction sur la longueur des preuves, en s'appuyant crucialement sur le fait que les règles de preuve de la théorie de base, en l'occurrence

85. Nous disons « nouveau sous-ensemble d'entiers naturels » au sens de non définissable dans le langage d'origine.

86. Bien qu'il soit très souvent fait référence à ce résultat, les démonstrations entièrement développées du principe de réflexion à partir d'une théorie tarskienne de la vérité sont à vrai dire assez rares dans la littérature. Le plupart des auteurs se contentent de donner la structure globale de la preuve accompagnée d'indications générales. Pour bien saisir le rôle des divers axiomes (compostionnels ou relevant de l'induction), un certain niveau de détail est néanmoins nécessaire. Nous nous appuyons ici sur la démonstration donnée dans Halbach (2014, § 8.6, p. 89-93).

PA, préservent la vérité.[87] Lorsque l'induction est donnée sous la forme d'un schéma d'axiomes, comme c'est le cas pour l'arithmétique de Peano du premier ordre, les règles de déduction sont en fait celles d'un système de preuve pour la logique du premier ordre. Pour montrer que ces règles préservent la vérité, les clauses compositionnelles sont

87. Pour quelques précisions supplémentaires : la preuve de l'énoncé (Ref) (*i.e.* de l'énoncé : « tous les théorèmes de PA sont vrais ») menée dans l'extension aléthique tarskienne procède typiquement en trois étapes :

1. On montre que tous les axiomes de la théorie sont vrais.
2. On montre que toutes les règles de déduction de la théorie préservent la vérité.
3. On raisonne par induction sur la longueur des preuves pour montrer que tous les théorèmes sont vrais.

La première étape (1.) emploie à la fois les axiomes $\{Tar\}$ et des instances de l'induction dans laquelle le prédicat de vérité apparaît (*i.e.* des instances de $SI(\mathcal{L}_{Vr})$). La seconde étape (2.) sert essentiellement à établir le pas d'induction qui sera employé dans la troisième étape (3.). En effet, le raisonnement sur la longueur des preuves se formalise dans PA^{+ind}_{Tar} au moyen d'un axiome d'induction de $SI(\mathcal{L}_{Vr})$. Voici une version de ce à quoi peut ressembler un tel axiome, tirée de Halbach (2014, chapitre 8, p. 92) :

$$\Big[\forall y(DemLong_{PA}(0,y) \rightarrow Vr(clu(y)) \wedge$$
$$\forall x\Big(\forall y(DemLong_{PA}(x,y) \rightarrow Vr(clu(y)) \rightarrow \forall y(DemLong_{PA}(Sx,y) \rightarrow Vr(clu(y))\Big)\Big] \rightarrow$$
$$\forall x \forall y(DemLong_{PA}(x,y) \rightarrow Vr(clu(y))$$

où

(a) $DemLong_{PA}(x,y)$ est une formule de $\mathcal{L}_{PA}$ exprimant le fait qu'il existe une preuve de (la formule codée par) y dont la longueur est inférieur à x.

En reprenant la notation du chapitre précédent, on peut par exemple prendre pour $DemLong(x,y) := \exists z(Prouv_{PA}(z,y) \wedge lg(z) \leq x)$ où $lg(z)$ désigne une fonction (récursive) qui à tout code d'une preuve associe la longueur de cette preuve, c'est-à-dire le nombre d'application(s) d'une règle de déduction

(b) et où $clu(y)$ désigne la clôture universelle de (la formule codée par) y. Le détour par les formules ouvertes et leurs clôtures universelles est nécessaires dans la mesure où la preuve d'un énoncé universel contiendra des sous-preuves portant sur des formules ouvertes.

Pour pouvoir conclure, il faut ensuite établir les deux conjoints qui forment l'antécédent de l'implication ci-dessus, à savoir

$$\forall y(DemLong_{PA}(0,y) \rightarrow Vr(clu(y))) \quad (i)$$
$$\text{et } \forall x\Big(\forall y(DemLong_{PA}(x,y) \rightarrow Vr(clu(y)) \rightarrow \forall y(DemLong_{PA}(Sx,y) \rightarrow Vr(clu(y))\Big) \quad (ii)$$

Autrement dit, il faut établir :

(i) le point de départ de l'induction, *i.e.* que les formules prouvables par une preuve de longueur 0 sont vraies. Bien entendu, les formules prouvables en une preuve de longueur 0 sont exactement les axiomes de la théorie. Le point de départ de l'induction est donc pris en charge par l'étape (1.)

(ii) le pas d'induction, *i.e.* le passage du fait que les formules prouvables par une preuve de longueur n sont vraies au fait que les formules prouvables par une preuve de longueur $n+1$ sont vraies. C'est ici que la préservation de la vérité par les règles de déduction de la théorie, obtenue à l'étape (2.) intervient.

Pour plus de détails nous renvoyons à nouveau à Halbach (2014, chapitre 8).

indispensables. En effet, les règles de déduction pour la logique du premier ordre sont intimement liées aux connecteurs logiques. Pour montrer qu'elles préservent la vérité, il faudra s'appuyer sur les liens qui relient la valeur sémantique [88] d'un énoncé complexe comprenant un connecteur logique en position principale aux valeurs sémantiques des sous-formules à partir desquelles il a été composé. Ceci apparaîtra plus clairement à travers un exemple : supposons que notre système de preuve soit donné sous la forme d'un système de déduction naturelle et qu'il contienne, entre autres règles, une règle d'introduction pour la conjonction :

$$\frac{A \qquad B}{A \wedge B}$$

Établir la correction de cette règle, c'est montrer que pour toute paire d'énoncés pouvant prendre la place des méta-variables A, B ci-dessus, lorsque les énoncés placés en position de prémisses sont vrais, l'énoncé placé en position de conclusion le sera également. Plus formellement, cela peut s'exprimer par l'énoncé du langage $\mathcal{L}_{Vr}$ suivant :

$$\forall x \forall y (En_{\mathcal{L}_{PA}}(x) \wedge En_{\mathcal{L}PA}(y) \rightarrow (Vr(x) \wedge Vr(y) \rightarrow Vr(x \dot{\wedge} y)$$

ce qui est quasiment une simple reformulation de la clause tarskienne pour la conjonction, à savoir

$$\forall x \forall y \left(En_{\mathcal{L}_T}(x \dot{\wedge} y) \rightarrow (Vr(x \dot{\wedge} y) \leftrightarrow (Vr(x) \wedge Vr(y))\right) [89]$$

Au coeur de la preuve de (Ref) discutée par Field, Shapiro et Ketland, on trouve donc en fait une version formalisée dans PA^{+ind}_{Tar} de ce qu'on nomme habituellement une preuve d'adéquation (ou de correction) [90] des règles de déduction de notre théorie de base. Et c'est justement la structure compositionnelle de la propriété de vérité, telle qu'elle se traduit dans les clauses tarskiennes, [91] qui permet de montrer que cette propriété de

88. Dans le cas qui nous concerne la valeur sémantique : « x est vrai ».
89. Où $T = PA$. Voyez le chapitre précédent pour une formulation explicite (possible) de $\{Tar\}$.
90. En anglais *soundness*.
91. Notez que l'axiomatisation donnée par la collection des **T**-équivalences suffit à dériver les instances de l'énoncé général « les règles préservent la vérité », mais ne suffit pas à obtenir l'énoncé général lui-même. Précisons ce que nous voulons dire par là en appuyant toujours sur notre exemple : pour φ et ψ *fixés*, la correction de

$$\frac{\varphi \qquad \psi}{\varphi \wedge \psi}$$

qui est une instance de la règle d'introduction pour la conjonction, s'obtient facilement à partir des seules **T**-équivalences de la manière suivante dans PA_d :

vérité, ou plus précisément son extension identifiée modulo un codage de la syntaxe à un sous-ensemble d'entiers, est close par application des règles de déduction de la théorie de base PA.

En dépit de la conservativité de $PA \cup \{Tar\}$ sur PA, il nous semble donc assez « audacieux » et pour tout dire peu plausible d'affirmer que la vérité (et en particulier les axiomes compositionnels) ne joue pas de rôle explicatif dans *cette* preuve [92] du principe de réflexion.

4.2.2.2 Contre-argument n°2 (Strollo (2013))

Pour le lecteur aux yeux duquel les considérations qui précèdent ne suffiraient pas, nous en venons à présent à un second argument avancé contre la réponse de Field. Ce second contre-argument repose sur des résultats relativement sophistiqués de théorie des modèles et plus particulièrement sur l'étude des classes de satisfaction. [93] Si les résultats techniques sur lesquels ce contre-argument s'appuie sont connus depuis relativement longtemps de la communauté des théoriciens des modèles, [94] l'idée de les prendre en considération pour discuter des mérites et des limites du déflationnisme remonte, semble-t-il, à McGee (2006) et Strollo (2013). [95]

$$\dfrac{\dfrac{}{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi}\text{T-éq} \quad \dfrac{}{Vr(\ulcorner\psi\urcorner) \leftrightarrow \psi}\text{T-éq}}{\dfrac{Vr(\ulcorner\varphi\urcorner) \wedge Vr(\ulcorner\psi\urcorner) \leftrightarrow \varphi \wedge \psi}{Vr(\ulcorner\varphi\urcorner) \wedge Vr(\ulcorner\psi\urcorner) \rightarrow Vr(\ulcorner\varphi \wedge \psi\urcorner)}\text{logique}}\text{T-éq + logique}$$

L'énoncé « $Vr(\ulcorner\varphi\urcorner) \wedge Vr(\ulcorner\psi\urcorner) \rightarrow Vr(\ulcorner\varphi \wedge \psi\urcorner)$ » est bien un énoncé exprimant la correction de l'instance de la règle pour la conjonction ci-dessus, au sens où il nous dit que si les prémisses *fixées* φ, ψ sont vraies alors la conclusion obtenue par application de la règle le sera également.

Les **T**-équivalences saisissent donc pour ainsi dire « instances par instances » la compositionnalité de la vérité et la correction des règles de déduction de la théorie de base. En revanche, elles sont trop faibles pour établir l'énoncé général de la correction.

Bien entendu, pour pouvoir mener à bien l'induction sur la longueur des preuves, c'est l'énoncé général exprimant la correction des règles qui est nécessaire. Et c'est pourquoi les axiomes compositonnels sont indispensables pour mener à bien la dérivation de (Ref).

92. Ce qui ne veut pas dire que cette preuve est particulièrement éclairante ou explicative. Ce que nous voulons dire ici, *contra* Field, c'est que, quoi que l'on pense par ailleurs de l'intérêt d'une preuve sémantique du schéma de réflexion puis de G ou de $Con(PA)$, il nous semble incongru de considérer que dans cette preuve le prédicat de vérité ne sert que comme outil expressif.

93. Voir ci-dessous pour une définition. L'étude des classes de satisfaction est aujourd'hui un domaine actif de recherche (pour une vue d'ensemble des principaux résultats, voir Kotlarski (1991) et Murawski (1997), pour un travail récent sur ce sujet, Enayat et Visser (2015)).

94. ou plus exactement de la communauté des mathématiciens qui étudient les modèles non standard de l'arithmétique

95. McGee (2006) introduit la conservativité modèle-théorique (voir ci-dessous) comme un critère qu'un déflationniste devrait souhaiter vérifier et propose une théorie compositionnelle plus faible que

Dans le contre-argument n°1 ci-dessus, l'idée centrale était d'examiner le rôle des axiomes compositionnels de la vérité dans la dérivation de (Ref) au sein de PA^{+ind}_{Tar}. On insistait sur leur indispensabilité en opposant à la conservativité de PA_{Tar} sur PA, la conservativité de PA_d sur PA^{+ind}_d. Ce dernier résultat de conservativité était censé fournir la preuve de l'insuffisance des seuls axiomes d'induction élargie. Dans le second contre-argument qui va nous occuper ici, le plan consiste plutôt à revenir plus en détails sur la conservativité PA_{Tar} sur PA, indépendamment du rôle jouée par l'élargissement de l'induction. En effet, s'il est exact que l'ajout à l'arithmétique de Peano des seules clauses tarskiennnes (sans élargir l'induction) ne permet pas de démontrer de nouveaux théorèmes du langage $\mathcal{L}_{PA}$, il ne faudrait peut-être pas en conclure pour autant que la notion de vérité ainsi axiomatisée n'a pas déjà un effet sur la structure possible du monde, qu'elle n'a pas déjà un impact sur les faits non sémantiques, même si cet impact ne se traduit pas sous la forme d'un nouveau théorème exprimé dans le langage de base. C'est en tout cas ce que défend Strollo (2013) : en dépit de la conservativité de l'extension aléthique obtenue, le passage de PA à PA_{Tar} revient bien à adopter des hypothèses portant sur une propriété substantielle et ayant des conséquences non strictement sémantiques, c'est-à-dire des conséquences sur la structure de modèles possibles de notre théorie. Simplement, pour percevoir ces effets il nous faut recourir à une notion plus fine que la notion de conservativité que nous avons jusqu'ici considérée.

Cette grille d'analyse au tamis plus fin que la conservativité habituelle s'appuie sur des éléments de théorie des modèles et s'articule autour de la notion d'expansion (ou d'enrichissement) d'un modèle. Rappelons que pour $\mathcal{L}$ un langage du premier ordre donné, une $\mathcal{L}$-structure d'interprétation $\mathfrak{M} := \langle M, \mathcal{I} \rangle$ est caractérisée par la donnée d'un ensemble non vide M, appelé domaine ou univers sous-jacent, et par une fonction d'interprétation $\mathcal{I}$ qui à chaque élément de la signature de $\mathcal{L}$ fait correspondre une interprétation sur M (*i.e.* une relation n-aire sur M pour chaque symbole de relation n-aire pris dans $\mathcal{L}$, une fonction n-aire sur M pour chaque symbole de fonction n-aire, un élément de M pour chaque symbole de constante)[96] Lorsqu'on considère une $\mathcal{L}$-structure $\mathfrak{M} := \langle M, \mathcal{I} \rangle$ donnée et qu'on enrichit le langage $\mathcal{L}$ en un langage $\mathcal{L}'$ en lui

$\{Tar\}$. Strollo (2013) emploie la conservativité modèle-théorique et des résultats portant sur les classes de satisfaction (que nous allons exposer en détails) pour contrer la stratégie de Field. C'est principalement cet argument dû à Strollo que nous reprenons ici.

96. Voyez n'importe quel manuel de logique. Habituellement, on note $R^{\mathfrak{M}}$, *etc.* l'interprétation du symbole R dans la structure $\mathfrak{M}$ de sorte que $\mathfrak{M}$ se note $\mathfrak{M} := \langle M, R^{\mathfrak{M}} \ldots, f^{\mathfrak{M}} \ \ldots, c^{\mathfrak{M}} \ \ldots \rangle$ où M est le domaine.

ajoutant de nouveaux symboles (de relations, fonctions ou constantes) on peut donc se demander s'il est possible de transformer cette $\mathcal{L}$-structure en une $\mathcal{L}'$-structure sans modifier ce que l'on avait déjà fixé, à savoir le domaine M et les interprétations sur M des symboles de $\mathcal{L}$.[97] Si tel est le cas on dit que $\mathfrak{M}$ peut être enrichie (ou étendue) en une $\mathcal{L}'$-structure. Avec cette notion d'expansion d'une $\mathcal{L}$-structure en une $\mathcal{L}'$-structure, on peut formuler une nouvelle notion de conservativité.

Pour comparaison, rappelons que nous avons déjà introduit deux notions de conservativité : la conservativité déductive (soient $T \subset T'$, T' est déductivement conservative sur T ssi pour tout énoncé φ de $\mathcal{L}_T$, $T' \vdash \varphi \Rightarrow T \vdash \varphi$) et la conservativité sémantique (mêmes hypothèses, T' est sémantiquement conservative sur T ssi pour tout énoncé φ de $\mathcal{L}_T$, $T' \models \varphi \Rightarrow T \models \varphi$). La complétude des systèmes de preuves pour la logique du premier ordre classique garantit que les deux notions coïncident et on obtient une notion double de conservativité. C'est cette double notion qui a été employée pour l'essentiel dans le débat provoqué par les arguments de Shapiro et Ketland. C'est notamment elle qui est invoquée par Field (1999). Pour la distinguer d'une autre notion de conservativtié que nous allons introduire nous l'appellerons désormais conservativité *syntaxique* dans ce qui suit. Il existe en effet un autre type de conservativité qu'on appelle parfois conservativité modèle-théorique et qui fut introduite par Craig et Vaught (1958) :

Définition 13. Conservativité modèle-théorique

Soient T une théorie exprimée dans un langage $\mathcal{L}_T$ et T' une théorie étendant T exprimée dans un langage $\mathcal{L}_{T'} \supseteq \mathcal{L}_T$. On dit que T' est modèle-théoriquement conservative sur T ssi toute $\mathcal{L}_T$-structure qui est modèle de T peut être enrichie en une $\mathcal{L}_{T'}$-structure qui est modèle de T'.

Autrement dit T' est modèle-théoriquement conservative sur T si et seulement si tout modèle de T peut être étendu en un modèle de T'. Si l'on se cantonne comme nous l'avons fait jusqu'ici à la logique du premier ordre, on peut se demander si cette autre notion de conservativité coïncide avec la notion de conservativité syntaxique. La réponse est non. En fait, cette notion de conservativité modèle-théorique est plus stricte que la notion habituelle de conservativité syntaxique : il est clair que si tout modèle de T peut être étendu en un modèle de T', alors T' sera syntaxiquement conservative sur

97. autrement dit, la question est de savoir si on peut « prolonger » la fonction d'interprétation $\mathcal{I}$ de $\mathcal{L}$ dans M en une fonction $\mathcal{I}'$ de $\mathcal{L}'$ dans M, sans toucher à M ni modifier $\mathcal{I}$.

T.[98] Mais l'inverse n'est pas toujours vrai : il existe des théories $T \subset T'$ respectivement exprimées dans $\mathcal{L}_T$ et $\mathcal{L}_{T'}$ telles que T' est une extension conservative de T, au sens syntaxique habituel, mais telles que T' n'est pas modèle-théoriquement conservative. Il peut parfois exister des théories $T \subset T'$ et des modèles de T qui ne peuvent être étendus à des modèles de T', alors même que T' est syntaxiquement conservative sur T.[99] Nous allons voir que c'est précisément le cas pour les théories $PA \subset PA_{Tar}$.

La notion de conservativité modèle-théorique est invoquée par Strollo (2013) comme un raffinement de la contrainte de conservativité développée par Shapiro et Ketland. Selon Strollo, c'est cette notion de conservativité modèle-théorique qui offre la meilleure façon de saisir la non substantialité de la vérité alléguée par les déflationniste :

> Quand nous nous intéressons à la métaphysique déflationniste de la vérité, cependant, ce qui devrait réellement importer est la question de savoir si *tous les modèles* de base peuvent être enrichis en des modèles de la théorie de base plus une théorie de la vérité. En fait, la question est de savoir si la vérité peut être ajoutée d'un manière telle que seule soit touché(e) (l'interprétation du) le langage et non pas l'univers sous-jacent. Si une extension ou une modification structurelle du domaine étaient requises, alors la vérité manifesterait des effets extralinguistiques. Elle affecterait les choses dont parle le langage et non pas seulement notre manière d'en parler. [...] je veux explicitement soutenir ici que la [conservativité modèle-théorique] [100] est la notion clé pour la non substantialité de la vérité déflationniste. Strollo (2013, p. 530).

Pour bien saisir ce qui se joue à travers l'argument de Strollo, le plus simple est peut-être de revenir à un exemple bien connu des logiciens et à une construction élémentaire de théorie des modèles. Considérez $\mathcal{L}_{PA} := \{0, +, ., s, <\}$ le langage de l'arithmétique muni d'un symbole de relation binaire dont l'interprétation attendue sera l'ordre habituel sur les entiers. Il est bien connu que les axiomes de Peano exprimés dans ce langage

98. Par l'absurde, supposons qu'il existe un $\mathcal{L}_T$-énoncé φ tel que $T' \models \varphi$ mais $T \not\models \varphi$. Alors, $T \cup \{\neg\varphi\}$ admet un modèle. Soit $\mathfrak{M} \models T \cup \{\neg\varphi\}$ un tel modèle. D'après la conservativité modèles-théorique, $\mathfrak{M}$ peut être enrichie en une $\mathcal{L}_{T'}$-structure $\mathfrak{M}'$ telle que $\mathfrak{M}' \models T'$. Mais alors $\mathfrak{M}' \models \varphi$, ce qui est impossible puisque $\mathfrak{M}'$ et un enrichissement de $\mathfrak{M}$ et que $\mathfrak{M} \models \neg\varphi$.

99. Les deux notions de conservativité syntaxique et modèle-théorique sont en fait liées de la manière suivante : T' exprimée dans $\mathcal{L}'$ est syntaxiquement conservative sur T exprimée dans $\mathcal{L}$ (avec $\mathcal{L} \subset \mathcal{L}'$ et $T \subset T'$) ssi tout modèle $\mathfrak{M} \models T$ se plonge élémentairement dans un modèle de T', *i.e.* pour tout modèle $\mathfrak{M} \models T$, il existe $\mathfrak{N} \models T'$ tel que $\mathfrak{M} \preccurlyeq \mathfrak{N} \restriction \mathcal{L}$.

100. En anglais Strollo parle d'expandability,*i.e.* capacité de ce qui peut être enrichi. On pourrait le rendre par des néologismes du type extensibilité, expansibilité, enrichissabilité...

admettent des modèles non standard. Parmi ce modèles non standard certains peuvent être obtenus de la manière suivante : soit T l'ensemble des énoncés de $\mathcal{L}_{PA}$ vrais dans le modèle standard, ce qu'habituellement on appelle la théorie complète de $\mathbb{N}$ et qu'on note $Th(\mathbb{N})$.[101] $T = Th(\mathbb{N})$ est bien une théorie exprimée dans $\mathcal{L}_{PA}$.[102] De plus, elle contient PA[103] et est trivialement complète pour $\mathcal{L}_{PA}$.[104] Soit c un nouveau symbole de constante qu'on adjoint à $\mathcal{L}_{PA}$ pour obtenir $\mathcal{L}' = \mathcal{L}_{PA} \cup \{c\}$. Soit T' l'extension de T (et de PA) suivante :

$$T' = Th(\mathbb{N}) \cup \{\underline{n} < c \mid n \in \mathbb{N}\}$$

Il est notoire que T' admet un modèle.[105] Bien évidemment, tout modèle de T' est un modèle de T et donc *a fortiori* de PA.[106] Par ailleurs, on vérifie sans aucune difficulté que T' est une extension conservative de T.[107]

Pour autant, peut-on considérer que le passage de T à T' s'est opéré sans qu'on ait ajouté de contenu « substantiel » à T ? La conservatité (syntaxique) de T' sur T doit-elle nous amener à conclure que la constante c telle qu'elle est caractérisée par les axiomes de T' n'a qu'un rôle purement expressif ? En acceptant les nouveaux axiomes

$$\{\underline{n} < c \mid n \in \mathbb{N}\}$$

exprimés dans le langage enrichi $\mathcal{L}_{PA} \cup \{c\}$, ne nous sommes-nous pas dotés au contraire d'une nouvelle notion substantielle pouvant jouer un rôle explicatif ? Malgré la conservativité syntaxique de T' sur T, il nous semble qu'on peut (et qu'on doit) raisonnablement douter de l'« innocence » de T' par rapport à T. En effet, tout modèle de T' sera ce qu'on appelle un modèle non-standard de PA (et de $Th(\mathbb{N})$). Du fait qu'il satisfait $\{\underline{n} < c \mid n \in \mathbb{N}\}$, tout modèle de T' sanctionnera l'existence d'un entier non-standard,

101. où $\mathbb{N} := \langle \omega, 0^{\mathbb{N}}, +^{\mathbb{N}}, .^{\mathbb{N}}, s^{\mathbb{N}}, <^{\mathbb{N}} \rangle$ est le modèle standard constitué des entiers naturels, où 0 désigne zéro et $+, ., s, <$ respectivement l'addition, la multiplication, la fonction successeur, et l'ordre habituels.
102. Même si bien sûr, cette théorie n'est pas récursivement axiomatisable.
103. car $\mathbb{N} \models PA$ et $Th(\mathbb{N}) = \{\varphi \in \mathcal{L}_{PA} \mid \mathbb{N} \models \varphi\}$. D'où, $PA \subset Th(\mathbb{N})$.
104. Pour tout énoncé φ de $\mathcal{L}_{PA}$, soit $\mathbb{N} \models \varphi$, soit $\mathbb{N} \models \neg\varphi$ et donc soit $\varphi \in T$ soit $\neg\varphi \in T$.
105. C'est un résultat élémentaire de logique, une application directe de la compacité. En fait, T' admet même une infinité de modèles (par Lowenheim-Skolem) et même un infinité non-dénombrable de modèles dénombrables ($2^{\aleph_0}$ pour être précis, ce qui se montre par des techniques plus complexes de théories des modèles, voir Kaye (1991)).
106. Plus exactement, le $\mathcal{L}_{PA}$-réduit de tout modèle de T' est un modèle de T et de PA.
107. Ceci découle immédiatement du fait que T est complète dans $\mathcal{L}$ et du fait que toute extension (cohérente) d'une théorie complète (cohérente) sera conservative : si $\mathcal{T}$ est complète et $\mathcal{T} \subset \mathcal{T}'$ alors pour $\varphi \in \mathcal{L}_{\mathcal{T}}$, $\mathcal{T}' \vdash \varphi \Rightarrow \mathcal{T} \vdash \varphi$ car si $\mathcal{T} \nvdash \varphi$ alors par complétude $\mathcal{T} \vdash \neg\varphi$, d'où $\mathcal{T}' \vdash \neg\varphi$, d'où, par cohérence de $\mathcal{T}'$, $\mathcal{T}' \nvdash \varphi$.

c'est-à-dire l'existence d'un objet différent de de tous les entiers naturels (et supérieur à ceux-ci pour la relation d'ordre dénotée par $<$). [108] En passant de T à T', on a donc accepté l'existence d'un tel objet. C'est, nous semble-t-il, faire une hypothèse aussi substantielle qu'une hypothèse peut l'être. Elle a évidemment un impact sur la structure des modèles possibles de notre théorie. L'une des conséquences les plus saisissantes de l'extension de T par T' est l'exclusion du modèle standard lui-même de la classe des modèles possibles de notre théorie : parmi les entiers naturels ne se trouve aucun objet susceptible de servir d'interprétation à c de manière à satisfaire $\{\underline{n} < c \mid n \in \mathbb{N}\}$. Autrement dit $\mathbb{N} := \langle \omega, 0^{\mathbb{N}}, +^{\mathbb{N}}, .^{\mathbb{N}}, s^{\mathbb{N}}, <^{\mathbb{N}} \rangle$ n'est pas modèle de T'. Cet impact ne se limite cependant pas à la seule existence d'un entier non-standard dénoté par c. On peut en effet montrer, par exemple, que tout modèle de T' possède une infinité d'éléments non-standard, que les entiers standard n'y sont pas définissables, que sa relation d'ordre n'est pas un bon ordre (contrairement à la relation habituelle sur les « vrais » entiers naturels), que tout sous-ensemble définissable non borné dans les entiers standard contient nécessairement un entier non-standard, [109] et tout un tas d'autres résultats fascinants qui constituent l'étude des modèles non-standard de l'arithmétique. [110] Dans les explications et les démonstrations de ces résultats les propriétés de l'objet dénoté par c telles qu'elles sont couchées dans les axiomes $\{\underline{n} < c \mid n \in \mathbb{N}\}$ jouent indéniablement un rôle central.

En résumé, par l'enrichissement du langage $\mathcal{L}_{PA}$ au moyen d'une nouvelle constante c et par l'extension de $Th(\mathbb{N})$ à T', nous nous sommes effectivement munis d'une nouvelle ressource expressive. En effet, les axiomes $\{\underline{n} < c \mid n \in \mathbb{N}\}$ nous assurent que l'interprétation de c sera distincte de celle de tous les termes du langage d'origine $\mathcal{L}_{PA}$; ils nous garantissent également que cet objet n'est pas définissable par une formule de $\mathcal{L}_{PA}$. Le gain en pouvoir expressif est donc net et évident. Mais ce n'est pas tout. En acceptant T', nous avons non seulement augmenté nos capacités expressives, mais nous avons également formulé —et adopté— ce qui semble bien être l'axiomatisation d'une notion substantielle pouvant jouer un rôle explicatif, à savoir en l'occurrence l'existence d'un entier non-standard. [111] Malgré cela, par construction T' est une extension syntaxique-

108. Plus rigoureusement : T' a pour conséquence l'existence d'un objet dénoté par c différent de tous ceux obtenus par un nombre fini d'applications de la fonction successeur appliqué à l'objet dénoté par 0.
109. C'est le lemme d'overspill.
110. Les modèles non standard de l'arithmétique forment un champ de recherche florissant de la logique mathématique contemporaine. Pour un tableau récent de ce domaine voir Kaye (1991) et Kossak et Schmerl (2006).
111. Notez que nous avons pris comme exemple le cas d'un élargissement de $\mathcal{L}_{PA}$ et de T au moyen d'une nouvelle constante et d'axiomes stipulant certaines propriétés fondamentales de (l'objet dénoté

ment conservative de T. On voit donc qu'ici cette forme de conservativité n'est pas assez fine pour détecter l'ajout (indéniable à nos yeux) de contenu qu'a constitué le passage de T à T'.

T' n'est en revanche pas modèle-théoriquement conservative sur T puisqu'il existe des $\mathcal{L}_{PA}$-structures, au premier rang desquelles le modèle standard $\mathbb{N}$ lui-même, qui sont des modèles de T mais qui ne peuvent être enrichies à un modèle de T'. Comme annoncé, on voit que la notion de conservativité modèle-théorique est plus stricte que la conservativité syntaxique habituelle. En termes modèles-théoriques, on peut reformuler les choses de la manière qui suit. Si pour tout $\mathcal{E}$ un ensemble d'énoncés, on note $Mod(\mathcal{E})$ la classe des modèles de $\mathcal{E}$, les inclusions ci-dessous sont strictes :

$$Mod(T') \overset{1}{\subsetneq} Mod(Th(\mathbb{N})) \overset{2}{\subsetneq} Mod(PA)$$

Ce qui n'est rien d'autre qu'une façon de réécrire qu'il existe une structure appartenant à $Mod(Th(\mathbb{N}))$ qui n'est pas dans $Mod(T')$ (et de même pour $Mod(PA)$ et $Mod(Th(\mathbb{N}))$).[112] Le caractère strict de ces inclusions reflète le fait que T' n'est pas *modèle-théoriquement* conservative sur $Th(\mathbb{N})$ qui n'est elle-même pas modèle-théoriquement conservative sur PA. Toutefois, seule la seconde inclusion se traduit par un résultat de non conservativité *syntaxique*. La raison en est la suivante : il existe une propriété exprimable par un énoncé de $\mathcal{L}_{PA}$ qui est satisfaite par toutes les structures qui sont des modèles de $Th(\mathbb{N}))$ mais qui n'est pas satisfaite par tous les modèles de

par) la constante. Ceci, pour rester le plus proches possible de la construction classique des modèles non-standard de PA, telle qu'on la trouve dans les manuels de logique. Mais, nous aurions tout aussi bien pu enrichir $\mathcal{L}_{PA}$ et T par un nouveau symbole de *prédicat unaire* gouverné par des axiomes étendant T. Une telle construction serait davantage similaire —du moins sur le plan de la morphologie et des catégories lexicales— au cas des extensions aléthiques. Pour illustration : soient $Vr(x)$ un nouveau symbole de prédicat unaire et $\mathcal{L}' = \mathcal{L}_{PA} \cup \{Vr\}$ un enrichissement de $\mathcal{L}_{PA}$. Soit T' la $\mathcal{L}'$-théorie étendant T de la manière suivante :

$$T' = Th(\mathbb{N}) \cup \{\neg Vr(\underline{n}) \mid n \in \mathbb{N}\} \cup \{\exists x Vr(x)\}.$$

Par compacité, T' est consistante. Mais tout modèle de T' sera nécessairement non standard puisqu'il devra contenir un objet satisfaisant Vr alors qu'aucun entier standard ne vérifie Vr. Pour d'autres exemples de ce type, voir McGee (2006, p. 105) et Strollo (2013, p. 523)

112. En toute rigueur, dans la mesure où $Mod(T')$ est une classe de $[\mathcal{L}_{PA} \cup \{c\}]$-structures, peut-être devrions nous plutôt écrire :

$$Mod(T') \restriction \mathcal{L}_{PA} \overset{1}{\subsetneq} Mod(Th(\mathbb{N})) \overset{2}{\subsetneq} Mod(PA)$$

PA.[113] À l'inverse, s'il existe bien des propriétés satisfaites par tous les modèles de T' sans être satisfaites par tous les modèles de $Th(\mathbb{N})$, comme par exemple « avoir un entier non-standard »[114], ou « ne pas être bien ordonnée par la relation dénotée par $<$ »[115], ... *etc.* ..., ***aucune*** de ces propriétés caractérisant ou distinguant $Mod(T')$ par rapport à $Mod(Th(\mathbb{N}))$ n'est exprimable par un énoncé de $\mathcal{L}_{PA}$.

Ainsi, on voit que le passage de $\mathcal{L}_{PA}$ à $\mathcal{L}'$ et de T à T' a bel et bien permis d'augmenter nos capacités expressives mais qu'il s'est également accompagné d'un effet sur la structure des modèles possibles de notre théorie. Nous sommes donc tentés de dire que les axiomes $\{\underline{n} < c \mid n \in \mathbb{N}\}$ ont un contenu substantiel pouvant jouer un rôle explicatif. Si l'ajout de ces axiomes ne débouche pas sur un nouveau théorème exprimé dans le langage d'origine, ce n'est pas parce que la notion ainsi formalisée (en l'occurrence l'existence d'un entier non standard) serait purement expressive et dénuée de tout contenu substantiel ou privée de rôle explicatif. Bien au contraire, c'est nous semble-t-il, tout simplement parce qu'ici pouvoir expressif et contenu substantiel ou capacité explicative sont indissociables. En passant de T à T' nous faisons bien une hypothèse lourde sur la structure de l'univers sous-jacent et cette hypothèse peut jouer un rôle explicatif. Mais précisément, et c'est là le point sur lequel nous voulons insister, cette hypothèse et les conséquences qu'elle renferme concernant la structure de l'univers ne sont pas exprimables dans le langage d'origine et ne peuvent donc pas se traduire sous la forme d'un énoncé de $\mathcal{L}_{PA}$ susceptible d'être prouvé.

Cet exemple avait pour but d'illustrer deux choses : d'un part, nous l'avons déjà dit, la conservativité modèle-théorique est strictement plus forte que la conservativité syntaxique. D'autre part, nous voulions aussi illustrer le fait que la conservativité syntaxique est parfois insuffisante pour établir l'absence de contenu substantiel ou de rôle explicatif potentiel d'une notion ; en particulier lorsque ce contenu n'est justement pas exprimable dans le langage d'origine. Il peut donc sembler raisonnable de renforcer la contrainte de conservativité (syntaxique) exprimée par Shapiro et Ketland en la remplaçant par une

113. Rien d'extraordinaire ici. Trivialement, par exemple, $Con(PA) \in Th(\mathbb{N})$ et par conséquent, $\forall \mathfrak{M} \in Mod(Th(\mathbb{N})), \mathfrak{M} \models Con(PA)$, tandis que $PA \cup \neg Con(PA)$ est consistante ce qui équivaut à $\exists \mathfrak{M}' \in Mod(PA)$ tel que $\mathfrak{M}' \models \neg Con(PA)$, *i.e.* $\mathfrak{M}' \not\models Con(PA)$.

114. Autrement dit être modèle de $\{\underline{n} < c \mid n \in \mathbb{N}\}$.

115. Plus exactement, dans tout modèle de T' on peut montrer qu'il existe une suite d'éléments infinie descendante

$$\ldots c_{n+1} < c_n < \cdots < c_0$$

contrainte de conservativité modèle-théorique. Pour reprendre une expression de McGee (2006, p. 106), il y a une manière « lourde » d'ajouter du contenu à une théorie : celle qui se traduit par la capacité de prouver un nouveau théorème (du langage d'origine) ; mais il existe également une façon plus « légère » de concevoir ce en quoi consiste ajouter un contenu à une théorie : selon cette conception, on ajoute quelque chose à une théorie lorsqu'on restreint la classe de ses modèles. Dans les deux cas, néanmoins, l'ajout n'est pas innocent et les notions nouvellement introduites ne sont pas purement expressives. Elles ont bien un impact sur la structure des modèles de la théorie. C'est pourquoi, selon Strollo (2013) et McGee (2006), c'est bien une contrainte de conservativité modèle-théorique qui doit s'appliquer à la vérité déflationniste.

Dans le cadre d'une analyse de la réponse de Field, le question est donc posée : PA_{Tar} est-elle modèle-théoriquement conservative sur PA ? Autrement dit, partant d'un modèle quelconque de PA, est-il toujours possible de l'étendre à un modèle de PA_{Tar} ? Plus précisément, si $\mathfrak{M} := \langle M, 0^{\mathfrak{M}}, +^{\mathfrak{M}}, .^{\mathfrak{M}}, s^{\mathfrak{M}}, <^{\mathfrak{M}}\rangle$ est une $\mathcal{L}_{PA}$-structure qui satisfait PA, existe-t-il toujours une sous-partie $V \subseteq M$ telle que si on interprète Vr par V on obtient une $\mathcal{L}_{PA} \cup \{Vr\}$-structure

$$\mathfrak{M}^* := \langle M, 0^{\mathfrak{M}}, +^{\mathfrak{M}}, .^{\mathfrak{M}}, s^{\mathfrak{M}}, <^{\mathfrak{M}}, Vr^{\mathfrak{M}^*} = V\rangle$$

telle que $\mathfrak{M}^* \models PA_{Tar}$ [116] ?

L'un des enseignement de l'étude des classes de satisfaction est justement que tel n'est pas le cas : PA_{Tar} *n*'est *pas* modèle-théoriquement conservative sur PA. Crucialement, cette non conservativité modèle-théorique s'obtient *sans* faire intervenir l'induction élargie au vocabulaire sémantique, ce qui va directement à l'encontre de la stratégie fieldienne pour « sauver » le déflationnisme.

Exposons à présent plus en détails les résultats invoqués par Strollo (2013), à commencer par quelques définitions.

Définition 14. Classes de satisfaction : *Soit $\mathfrak{M}$ un modèle de PA de domaine M. Une sous-partie $S \subseteq M$ est appelée une classe de satisfaction (complète) pour $\mathfrak{M}$ si et seulement si $\langle \mathfrak{M}, S\rangle \models PA_{Tar}$.* [117]

116. Remarquez que l'on exige *pas* que $Vr^{\mathfrak{M}^*}$ satisfasse l'induction, *i.e.* que $\mathfrak{M}^* \models PA^{+ind}_{Tar}$.

117. Ici, on pourrait plutôt parler de classe de vérité, puisqu'il s'agit de trouver une sous-partie de M pouvant servir d'extension à Vr. La notion originelle (ainsi que l'appellation aujourd'hui bien implantée) concernait une sous-partie de M^2 interprétant une relation binaire $S(x, y)$ vérifiant des axiomes tarskiens pour la satisfaction (x sera le code d'une formule, y le code d'une séquence d'objets). Pour plus de détails

Définition 15. TYPES : *Soit $\mathcal{L}$ un langage et $\mathfrak{M}$ une $\mathcal{L}$-structure. Un type $p(\vec{x})$ sur $\mathfrak{M}$ est un ensemble de formules de $\mathcal{L}$ à paramètres dans M ayant toutes exactement $\vec{x}$ comme variables libres, et tel que pour tout sous-ensemble fini $\Gamma(\vec{x})$ de $p(\vec{x})$, il existe une suite d'éléments $\vec{m} \in M$ telle que $\mathfrak{M} \models \varphi(\vec{m})$ pour toute $\varphi(\vec{x}) \in \Gamma(\vec{x})$ (on dit que $p(\vec{x})$ est finiment satisfaisable dans $\mathfrak{M}$).*

Si l'ensemble des codes des formules de $p(\vec{x})$ est récursif, on dit que $p(\vec{x})$ est un type récursif sur $\mathfrak{M}$.

Un type $p(\vec{x})$ sur $\mathfrak{M}$ est réalisé dans $\mathfrak{M}$ si et seulement si il existe une suite d'éléments $\vec{m} \in M$ telle que $\mathfrak{M} \models \varphi(\vec{m})$ pour toute $\varphi(\vec{x}) \in p(\vec{x})$.

Définition 16. SATURATION RÉCURSIVE : *Un modèle $\mathfrak{M}$ de PA est dit récursivement saturé si et seulement si $\mathfrak{M}$ réalise tous ses types récursifs.*

Théorème 17. *(*LACHLAN *(1981))*

Soit $\mathfrak{M}$ un modèle non standard de PA. Si $\mathfrak{M}$ possède une classe de satisfaction alors $\mathfrak{M}$ est récursivement saturé.

Lorsqu'on se restreint à des structures aux domaines dénombrables, on a une forme de réciproque :

Théorème 18. *(*KOTLARSKI, KRAJEWSKI *et* LACHLAN *(1981))*

Soit $\mathfrak{M}$ un modèle non standard de PA dénombrable[118] *et récursivement saturé. Alors, $\mathfrak{M}$ admet une classe de satisfaction.*

Les démonstrations de ces deux théorèmes sont difficiles et nécessitent des techniques avancées de théorie des modèles. Nous ne les donnerons pas ici.[119] Ces théorèmes ont toutefois des conséquences remarquables pour la discussion qui nous occupe.

voir KAYE (1991, chapitre 15, p. 224-225). Mentionnons également que les spécialistes distinguent les classes de satisfactions complètes par opposition aux classes de satisfaction partielles (où le degré de complexité des formules vérifiant les axiomes sémantiques est borné par un entier non standard) et les classes de satisfactions inductives ou non selon que S satisfait ou non l'indution, *i.e.* selon que $\langle M, S\rangle \models PA^{+ind}_{Tar}$ ou seulement $\langle M, S\rangle \models PA_{Tar}$. Ici, nous nous restreignons aux classes de satisfaction complètes sans exiger qu'elles soient inductives (mais sans l'exclure non plus).

118. Signalons, que la restriction aux modèles dénombrables est indispensable. SMITH (1984, 1989) a montré qu'il existe des modèles indénombrables qui, quoique récursivement saturés, n'admettent pas de classe de satisfaction (voir KOTLARSKI (1991)).

119. HALBACH (2014, chapitre 8, section 8.4) contient une démonstration du théorème de Lachlan, reprise en partie de KAYE (1991). Halbach cite également le théorème de Kotlarski *et alii*, mais sans en donner de démonstration. Pour une démonstration, voir KAYE (1991) ainsi que ENGSTRÖM (2002).

Premièrement, couplé à un autre résultat bien connu de théorie des modèles, le second théorème ci-dessus nous donne une démonstration modèle-théorique [120] de la conservativité *syntaxique* de PA_{Tar} sur PA. Sachant que,

Lemme. *Tout modèle dénombrable $\mathfrak{M}$ de PA possède une extension élémentaire dénombrable récursivement saturée,* [121]

il suit facilement que pour tout énoncé $\varphi \in \mathcal{L}_{PA}$, si $PA_{Tar} \vdash \varphi$ alors $PA \vdash \varphi$. En effet, supposons que $PA \nvdash \varphi$, alors $PA \cup \{\neg\varphi\}$ est consistant. Par Lowenheim-Skolem, il existe $\mathfrak{M} \models PA \cup \{\neg\varphi\}$ dénombrable. D'après le lemme, il existe $\mathfrak{N}$ dénomblable et récursivement saturé tel que $\mathfrak{M} \preccurlyeq \mathfrak{N}$. Par équivalence élémentaire, $\mathfrak{N} \models PA \cup \{\neg\varphi\}$. Par le théorème 18 ci-dessus, $\mathfrak{N} \models PA_{Tar}$, et comme $\mathfrak{N} \models \neg\varphi$, il suit que $PA_{Tar} \nvdash \varphi$.

En second lieu, le théorème de Lachlan 17 est quant à lui au centre de l'argument de Strollo. Adossé à un autre résultat de théorie des modèles :

Lemme. *Il existe des modèles non standard de PA non récursivement saturés.* [122]

ce théorème a pour corollaire direct le résultat suivant :

Corollaire. *Il existe des modèles de PA qui ne possèdent pas de classes de satisfaction.*

Ce qui revient à dire qu'il existe des modèles de PA qui ne sont pas des modèles de PA_{Tar}, c'est-à-dire que

Propriété. PA_{Tar} **n'**est **pas** *modèle-théoriquement conservative sur PA.*

En résumé, dès lors qu'on traite le schéma d'induction comme *liste* et qu'on ne permet pas aux formules contenant « Vr » d'apparaitre dans une induction, lorsqu'on adjoint les clauses compositionnelles tarskiennes $\{Tar\}$ à l'arithmétique de Peano, on obtient bien une extension syntaxiquement conservative. Cependant, l'extension ainsi obtenue, PA_{Tar}, n'est pas modèle-théoriqueemnt conservative sur PA. Autrement dit,

120. Historiquement, c'est la première démonstration du fait que PA_{Tar} étend conservativement (au sens syntaxique) PA a avoir été donnée. Elle précède la démonstration par élimination des coupure formulée par Halbach (1999a) et corrigée par Leigh (2013). Enayat et Visser (2015) ont proposé une autre preuve modèle-théorique améliorant celle de Kotlarski, Krajewski et Lachlan (1981), et ils ont également montré comment leur preuve pouvait se formaliser dans l'arithmétique primitive récursive.

121. Voir Kaye (1991, Proposition 11.4, p. 148)

122. Il en existe même des dénombrables. Pour une démonstration en bonne et due forme, nous renvoyons à nouveau à Kaye (1991). Notez que les modèles premiers, en particulier, ne sont pas récursivement saturés (Kaye, 1991, chapitre 8). Halbach (2014, Lemme 8.16, p. 75) contient également une construction explicite d'un modèle non standard de PA non récursivement saturé.

le simple fait de postuler les clauses récursives à la Tarski pour la vérité a déjà un impact sur la structure des modèles possibles, bien que ceci ne se traduise paspar la possibilité de dériver dans la théorie étendue de nouveaux théorèmes exprimés dans le langage de l'arithmétique. Pour Strollo (2013), cela signifie que ce que Field appelle les « axiomes essentiels de la vérité » a déjà un contenu substantiel pouvant jouer un rôle dans nos explications. [123] La stratégie de Field (1999) qui entendait ne voir dans le prédicat de vérité axiomatisé par $\{Tar\}$ qu'un auxiliaire expressif permettant de renforcer l'induction, mais sans responsabilité directe [124] dans l'augmentation de nos capacités de preuves est donc bloquée. Bien entendu, pour détecter correctement le contenu de $\{Tar\}$ (hors induction élargie), il a fallu recourir à des outils de théorie des modèles permettant une mesure plus sophistiquée, plus fine et plus stricte que la simple conservativité syntaxique. Au demeurant, l'un des avantages de la conservativité modèle-théorique est d'être moins soumise aux limitations et faiblesses expressives du langage de base, puisqu'ici on raisonne directement sur les (propriétés des) structures et non pas sur les énoncés du langage de base que ces structures vérifient (ce qui revient à se limiter aux seules propriétés exprimables dans ce langage). Cela s'est traduit par une reformulation et même par un renforcement de la contrainte de conservativité à partir de la notion d'expansion d'un modèle débouchant sur la conservativité modèle-théorique. Selon Strollo, toutefois, cette reformulation respecte l'esprit si ce n'est la lettre de l'argument de Keltand et Shapiro. [125] Voilà pour le contre-argument n°2, dû à Strollo (2013).

On peut toutefois en poursuivre le développement, notamment en approfondissant le parallèle avec la construction de modèles non standard de l'arithmétique.

De la même manière que dans le cas de la construction des modèles non standard de l'arithmétique complète $Th(\mathbb{N})$, si l'on reprend la notation de théorie des modèles introduite précédemment, on peut remarquer que les inclusions suivantes entre classes de modèles sont strictes :

$$Mod(PA^{+ind}_{Tar}) \overset{1}{\subsetneq} Mod(PA_{Tar}) \overset{2}{\subsetneq} Mod(PA)$$

123. Peut-être en collaboration avec d'autres axiomes ou d'autres principes de preuve.
124. au sens où ce prédicat désignerait une propriété substantielle pouvant jouer un rôle explicatif
125. D'ailleurs, notons que Shapiro lui-même fait allusion à la possibilité d'étendre tous les modèles de base à des modèles d'une extension aléthique, mais sans la relier explicitemnt à la notion précise de conservativité modèle-théorique (voyez Shapiro (1998b, p. 497 ?)).

Là encore, le caractère strict de ces inclusions traduit le fait que PA_{Tar} (respectivement PA^{+ind}_{Tar}) n'est pas modèle-théoriquement conservative sur PA (respectivement sur PA_{Tar} [126])

Pour autant, comme dans le cas de la construction des entiers non standard, seule la première inclusion ci-dessus s'accompagne d'un résultat de non conservativité syntaxique. [127] La raison profonde de cette différence est la même que précédemment : il n'existe pas d'énoncé exprimable dans le langage de base $\mathcal{L}_{PA}$ qui permet de distinguer la classe des modèles de PA_{Tar} de celles des modèles de PA, quoique la première soit strictement plus petite et contenue dans la seconde. Faut-il en conclure que l'élargissement de PA à PA_{Tar} est « innocent » et que le prédicat de vérité ainsi axiomatisé est non substantiel et dénué de pouvoir explicatif ? Nous ne le croyons pas ; comme dans le cas des entiers non standard, la conservativité *syntaxique* n'est pas due ici à l'absence de contenu, à la maigreur [128] métaphysique ou explicative de la vérité. Elle provient tout simplement, nous semble-t-il, du fait qu'ici encore, à l'instar de ce que nous avons noté dans le cas de l'extension $Th(\mathbb{N}) \subset Th(\mathbb{N}) \cup \{\underline{n} < c \mid n \in \mathbb{N}\}$, pouvoir expressif et contenu substantiel ou capacité explicative sont inséparables : la propriété axiomatisée par $\{Tar\}$ a bien un contenu substantiel, ce dont témoigne la non-conservativité modèle-théorique. Mais cette propriété substantielle, celle qui est partagée par toutes les structures de la classe $Mod(PA_{Tar})$ n'est justement pas exprimable dans $\mathcal{L}_{PA}$. Et, c'est pour cette raison que la non conservativtié modèle-théorique ne s'accompagne pas ici d'un nouveau théorème exprimé dans le langage de base, autrement dit d'un phénomène de non conservativité *syntaxique*. On peut d'ailleurs rendre le parallèle avec le cas de l'extension de $Th(\mathbb{N})$ par $\{\underline{n} < c \mid n \in \mathbb{N}\}$ plus évident encore en considérant le cas suivant : partons à nouveau de la théorie $Th(\mathbb{N})$ de l'arithmétique complète dans $\mathcal{L}_{PA}$. On a le résultat suivant concernant les classes de satisfaction :

Lemme. *Toute extension consistante de PA admet un modèle non récursivement saturé.* [129]

Or, il va de soi que $Th(\mathbb{N})$ est une extension consistante de PA. Donc, $Th(\mathbb{N})$ admet des modèles non recursivement saturés. De tels modèles seront évidemment des modèles de PA (trivialement, puisque $Mod(Th(\mathbb{N})) \subsetneq Mod(PA)$). Le théorème de Lachlan a

126. et donc *afortiori* PA^{+ind}_{Tar} n'est pas modèle-théoriquement conservative sur PA.
127. Par exemple, $PA^{+ind}_{Tar} \vdash G$ alors que $PA_{Tar} \nvdash G$ (et bien sûr $PA \nvdash G$).
128. pour reprendre la terminologie de Field (1999) et Shapiro (1998b).
129. *Cf.* Halbach (2014, p.75), par exemple, pour une démonstration.

donc pour conséquence immédiate qu'il existe des modèles de $Th(\mathbb{N})$ qui n'admettent pas de classes de satisfaction. Autrement dit, il existe des modèles de $Th(\mathbb{N})$ qui ne peuvent se prolonger en des modèles de $Th(\mathbb{N}) \cup \{Tar\}$. Ainsi, l'inclusion

$$Mod(Th(\mathbb{N}) \cup \{Tar\}) \subsetneq Mod(Th(\mathbb{N}))$$

est stricte ; autrement dit encore, $Th(\mathbb{N}) \cup \{Tar\}$ n'est pas modèle-théoriquement conservative sur $Th(\mathbb{N})$; elle est en revanche bien évidemment *syntaxiquement* conservative sur $Th(\mathbb{N})$ (ne serait-ce que parce que $Th(\mathbb{N})$ est complète pour $\mathcal{L}_{PA}$ [130])

Il semble donc que la situation concernant l'extension $Th(\mathbb{N}) \cup \{Tar\}$ par rapport à $Th(\mathbb{N})$ soit rigoureusement la même que celle de l'extension $Th(\mathbb{N}) \cup \{\underline{n} < c \mid n \in \mathbb{N}\}$. De même qu'il nous semble pour le moins contre-intuitif d'affirmer que les axiomes d'existence d'un entier non standard formalisent une notion non substantielle, ou non explicative, ou purement expressive… au motif que $Th(\mathbb{N}) \cup \{\underline{n} < c \mid n \in \mathbb{N}\}$ est *syntaxiquement* conservative sur $Th(\mathbb{N})$, de même nous ne voyons aucune raison de suivre Field pour attribuer ces qualités (ou ces absences de qualités) au prédicat de vérité axiomatisé par les clauses tarskiennes, avec pour seule justification la conservativité *syntaxique* de PA_{Tar} sur PA. Sur ce point, nous nous rangeons donc aux côtés de Strollo (2013).

Pour finir sur ce contre-argument n°2, nous voudrions évoquer également une autre conséquence découlant de l'analyse des classes de satisfaction et ayant une importance pour le débat sur le déflationnisme. La classe de structures $Mod(PA_{Tar})$ est évidemment (infiniment) axiomatisable dans $\mathcal{L}_{PA} \cup \{Vr\}$. [131] Mais l'une des conséquences des résultats que nous avons mentionnés dans cette sous-section est que la classe des $\mathcal{L}_{PA}$-structures constituée des restrictions au langage $\mathcal{L}_{PA}$ des structures comprises dans $Mod(PA_{Tar})$, ce qu'on pourrait noter

$$Mod(PA_{Tar}) \restriction \mathcal{L}_{PA} = \{\mathfrak{M} \restriction \mathcal{L}_{PA} \mid \mathfrak{M} \in Mod(PA_{Tar})\}$$

n'est pas (même infiniment) axiomatisable dans $\mathcal{L}_{PA}$. En d'autres termes, il n'existe

130. Aucune chance en effet de trouver un nouveau théorème de $\mathcal{L}_{PA}$ non démontré par $Th(\mathbb{N})$, sauf à se placer dans une extension incohérente.
131. Et pour cause, elle est bien évidemment axiomatisée par PA_Tar.

aucun ensemble (même éventuellement infini) Γ d'énoncés de $\mathcal{L}_{PA}$ tel que

$$Mod(\Gamma) = Mod(PA_{Tar})$$ [132]

En effet, supposons qu'un tel ensemble Γ vérifiant $Mod(\Gamma) = Mod(PA_{Tar})$ existe. Alors, $\Gamma^{\vdash} = \{PA_{Tar}\}^{\vdash} \cap \mathcal{L}_{PA}$. [133] Mais dans ce cas, $PA^{\vdash} \subset \Gamma^{\vdash} = \{PA_{Tar}\}^{\vdash} = PA^{\vdash}$ (par conservativité syntaxique de PA_{Tar} sur PA dans $\mathcal{L}_{PA}$). D'où, $\Gamma^{\vdash} = PA^{\vdash}$, et $Mod(\Gamma) = Mod(PA)$. Or $Mod(PA) \neq Mod(PA_{Tar}) \upharpoonright \mathcal{L}_{PA}$. Contradiction. Rien de renversant jusque là. Mais remarquez que si on s'intéresse à présent aux propriétés exprimables par des conjonction *infinies* d'énoncés de $\mathcal{L}_{PA}$ et qu'on interprète ces « formules » infinies de la manière qui semble s'imposer naturellement :
Pour tout ensemble Γ d'énoncés

$$\mathfrak{M} \models \bigwedge_{\varphi \in \Gamma} \varphi \text{ ssi } \forall \varphi \in \Gamma, \mathfrak{M} \models \varphi \text{ ssi } \mathfrak{M} \models \{\varphi \mid \varphi \in \Gamma\}$$

il s'en suit que $Mod(PA_{Tar})$, la classe des structures axiomatisée par PA_{Tar} ne peut être axiomatisée par aucune conjonction infinie d'énoncés de $\mathcal{L}_{PA}$:

$$\forall \Gamma \subset En(\mathcal{L}_{PA}), Mod(PA_{Tar}) \neq Mod(\bigwedge_{\varphi \in \Gamma} \varphi)$$
$$i.e.\ \forall \Gamma \subset En(\mathcal{L}_{PA}), Mod(PA_{Tar}) \neq Mod(\{\varphi \mid \varphi \in \Gamma\ \})$$

Ce qui revient à dire que la propriété exprimée par PA_{Tar}, caractéristique des structures de la classe $Mod(PA_{Tar})$ n'est pas exprimable par une conjonction infinie d'énoncés de $\mathcal{L}_{PA}$. Voici un bien curieux résultat si on le confronte à la thèse déflationniste selon laquelle le prédicat de vérité n'est qu'un outil purement expressif permettant d'exprimer des généralisations, lesquelles généralisations sont habituellement conçues comme des conjonction infinies. [134] [135]

132. ou plutôt, à strictement parler, tel que

$$Mod(\Gamma) = Mod(PA_{Tar}) \upharpoonright \mathcal{L}_{PA}$$

133. où $X^{\vdash}$ désigne la clôture déductive de X par les règles de déduction de la logique du premier ordre classique.

134. Voyez les exemples typiques de montée sémantique que l'on retrouve chez à peu près tous les auteurs déflationnistes.

135. Pour être exhaustifs sur l'emploi de la conservativité modèle-théorique dans les discussions sur le déflationnisme, signalons encore deux résultats importants que nous discuterons pas plus avant ici.

4.2.2.3 Contre-argument n°3 (Halbach (2014, 1999a, 2001b), Shapiro (2003))

Le troisième et dernier contre-argument que nous allons exposer ici s'éloigne quelque peu d'une analyse à la lettre de la réponse de Field. Ce contre-argument ne consiste pas à disséquer finement les résultats de conservativité (ou de non conservativité) syntaxique ou modèle-théorique de PA^{+ind}_{Tar} ou de PA_{Tar}, ni à se demander quel est le rôle joué par tel ou tel axiome dans la dérivation d'un nouveau théorème. Au lieu de cela, la démarche qui motive cet ultime contre-argument s'attache à tenter d'évaluer précisément le contenu supplémentaire de PA^{+ind}_{Tar} par rapport à PA. Ce faisant, on obtiendra ce qu'Halbach (2014) qualifie de meilleur argument contre l'idée que les axiomes pour la vérité sont conservatifs sur la théorie de base. Selon ses propres mots :

> De toute façon, le théorème 8.33 [établissant la non conservativité de PA^{+ind}_{Tar} sur PA en donnant une dérivation de $Con(PA)$] et autres phénomènes gödeliens ne fournissent pas les considérations les plus fortes contre les conceptions de la vérité qui regardent les axiomes pour la vérité comme étant conservatifs sur la théorie de base. Je pense qu'une comparaison avec la quantification du second ordre est plus révélatrice. [...]
>
> Le résultat est également plus instructif [...] puisque le système du second ordre ACA auquel $[PA^{+ind}_{Tar}]$ sera relié est bien compris et beaucoup plus fort que le simple énoncé de la consistance de l'arithmétique de Peano. Ainsi, c'est la comparaison avec ACA plutôt que le théorème 8.33 qui révèle toute l'étendue de la force du prédicat de vérité compositionnel de $[PA^{+ind}_{Tar}]$. Halbach (2014, p. 94)

Dans son article de 2006, Vann McGee propose une axiomatisation pour « Vr » qui, lorsqu'on n'élargit pas l'induction étend conservativement (au sens modèle-théorique) PA et qui, une fois l'induction étendue devient équivalente à PA^{+ind}_{Tar}(et est donc réflexive). Pour McGee, c'est cette axiomatisation que devrait adopter un déflationniste souhaitant poursuivre la stratégie ébauchée par Field. En effet, elle semble effectivement répondre à la double exigence de neutralité pré-élargissement de l'induction, et de capacité à prouver (Ref) post-élargissement de l'induction. Néanmoins, l'axiomatisation proposée par McGee souffre d'autres problèmes. En outre, si cette axiomatisation est peut-être une réponse satisfaisante au contre-argument n°2, elle reste toutefois soumise au contre-argument n°1 ci-dessus ainsi qu'au contre-argument n °3 (voir ci-desssous).

L'autre résultat surprenant qu'il nous faut indiquer et qui est déjà mentionné par Strollo (2013) est le suivant : PA_d n'est pas modèle-théoriquement conservative sur PA. Ainsi, si la contrainte de conservativité modèle-théorique traduit correctement les allégations de non substantialité et d'absence de pouvoir explicatif prononcées par les déflationnistes au sujet de la vérité, alors l'extension aléthique obtenue en ajoutant les seules **T**-équivalences et en élargissant l'induction est déjà trop forte pour être acceptable. Ce résultat étonnant est lui aussi une conséquence inattendue de l'étude modèle-théorique des classes de satisfaction. Sa démonstration est attribuée par Strollo à Fredrik Engström (voyez Strollo (2013)).

Notez que pour Halbach, à qui nous empruntons les considérations de cette sous-section, la comparaison de PA^{+ind}_{Tar} avec ACA ne vaut pas argument contre le déflationnisme, dans la mesure où selon Halbach, [136] le déflationnisme n'est pas tenu par une contrainte de conservativité ni plus généralement par l'idée que l'ajout de la notion de vérité ne doit pas s'accompagner de conséquences substantielles. [137] Pour autant, l'analyse d'Halbach (2014, § 8.6, p. 88-102), que nous reprenons dans ce qui suit, semble clairement en contradiction avec la stratégie défendue par Field (1999). Cette analyse et les résultats techiniques sur lesquels elle s'appuie est d'ailleurs reprise par Shapiro (2003) contre Field. C'est pourquoi nous la classons parmi les contre-arguments.

Pour évaluer précisément le contenu mathématique de PA^{+ind}_{Tar} par rapport à celui de PA, Halbach peut s'appuyer sur des résultats classiques de théorie réductive de la preuve. [138] À bien des égards, le programme de recherche de la théorie réductive de la preuve, qui culmine avec les mathématiques inversées, peut être considéré comme l'héritier contemporain des travaux de Hilbert sur les fondations des mathématiques. [139] La question principale qui sous-tend ce domaine consiste à déterminer précisément quelles sont les hypothèses ensemblistes minimalement nécessaires pour pouvoir développer telle ou telle partie des mathématiques ordinaires. [140] Le cadre technique permettant l'étude minutieuse de cette question est généralement $\mathcal{Z}_2$ et ses sous-fragments. Le système axiomatique $\mathcal{Z}_2$, parfois appelé arithmétique du second ordre [141] ou analyse classique, [142] est

136. contrairement à Field ?
137. voyez la section 4.1.1 au début de ce chapitre.
138. Pour une introduction à ce domaine de recherche mathématique, voyez Simpson (2009). Le premier chapitre, en particulier, offre une vue synthétique à la fois des motivations philosophiques, du cadre de travail employé et des principaux résultats techniques obtenus.
139. Pour l'anecdote, la filiation avec Hilbert est explicitement revendiquée dans Simpson (2009, chapitre 1, p. 6)
140. Mathématiques « ordinaires » par opposition à la théorie abstraite des ensembles issues des travaux de Cantor, car bien entendu cette question n'a d'intérêt que pour les mathématiques qui ne sont pas, ou pas directement, de la théorie des ensembles. La délimitation rigoureuse des mathématiques ordinaires, *i.e.* des mathématiques qui se sont développées antérieurement ou indépendamment de la théorie abstraite des ensembles, n'est pas toujours évidente (*cf.* Simpson (2009, chapitre 1) pour une discussion). Néanmoins, il est clair que si tout un chacun reconnaîtra l'importance des grands cardinaux pour l'étude axiomatique des ensembles, on peut se demander si (ou dans quelle mesure, à quel moment) de tels axiomes sont indispensables pour le développement, par exemple, de l'analyse numérique ou du calcul différentiel, *etc.* Il y a donc bien un sens à se demander quelles parties des mathématiques « non-ensemblistes » peuvent être reconstruites dans des extensions plus ou moins fortes de l'arithmétique.
141. ce nom est un peu trompeur dans la mesure où dans les approches contemporaines que nous discutons ici, ce système est en réalité muni d'une sémantique du premier ordre à deux sortes de variables.
142. Puisque modulo un codage des réels comme ensembles infinis d'entiers, une bonne partie de l'analyse classique peut y être reconstruite.

une extension stricte de PA tant sur le plan du langage que des axiomes.[143] En voici une présentation rapide inspirée de Simpson (2009) :

Définition. Arithmétique du second ordre ($\mathcal{Z}_2$)

Langage : Le langage $\mathcal{L}_2$ de l'arithmétique du second ordre est un langage du *premier ordre* à deux sortes de variables. On peut le voir comme un enrichissement de $\mathcal{L}_{PA}$. Outre les variables individuelles habituelles $x, y, z, \ldots, n, m, \ldots$, dites numériques (ou de la première sorte, ou du premier ordre) censées désigner des entiers naturels, $\mathcal{L}_2$ contient également des variables d'une seconde sorte $X, Y, Z, \ldots$ appelées variables d'ensemble (ou du second ordre) et censées désigner des ensembles d'entiers.

Comme symboles non logiques $\mathcal{L}_2$ contient outre les symboles de $\mathcal{L}_{PA}$ un symbole binaire d'appartenance $\in$. Ainsi la signature de $\mathcal{L}_2$ est : $\{0, +, ., s, <, \in\}$.

Les termes numériques t_i formés à partir de 0, de la fonction unaire s et des opérations binaires $+$ et $.$ sont les mêmes que ceux de $\mathcal{L}_{PA}$.

Les formules atomiques sont les mêmes que celles de $\mathcal{L}_{PA}$, *i.e.* du type $t_i = t_j$, $t_i < t_j$, auxquelles s'ajoutent les formules du type $t_i \in X$ où t_i est un terme numérique et X une variable de la seconde sorte.

Les règles de formation des formules complexes et de quantification sont les règles habituelles pour un langage du premier ordre à deux sortes.[144]

Axiomes : Les axiomes de $\mathcal{Z}_2$, l'arithmétique du second ordre, sont les clôtures universelles des formules suivantes :

i) [Axiomes de base] : ce sont les mêmes que les axiomes (du premier ordre) de

143. L'emploi de $\mathcal{Z}_2$ comme cadre fondationnel des mathématiques hors théorie des ensembles remonte à Hilbert et Bernays (1939). Signalons que $\mathcal{Z}_2$ quoiqu'extrêmement puissant comme système axiomatique au point de pouvoir servir de cadre à presque toutes les mathématiques classiques est néanmoins bien plus faible que ZF.

144. On a donc deux sortes de quantifications existentielles : $\exists x$ et $\exists X$, et deux universelles : $\forall x$ et $\forall X$.

PA hors induction. [145]

ii) [Axiome d'induction] : du fait de la présence de variables d'ensemble, l'induction peut se donner sous la forme d'un unique axiome :

$$(0 \in X \wedge \forall n(n \in X \rightarrow s(n) \in X)) \rightarrow \forall n(n \in X)$$

ii') [Schéma d'induction] il est néanmoins également possible de traiter l'induction sous la forme d'un schéma :

$$\Big(\varphi(0) \wedge \forall n(\varphi(n) \rightarrow \varphi(s(n)))\Big) \rightarrow \forall n\varphi(n)$$

où φ est une formule quelconque de $\mathcal{L}_2$.

iii) [Schéma de compréhension] :

$$\exists X \forall n(n \in X \leftrightarrow \varphi(n))$$

où $\varphi(n)$ est une formule quelconque de $\mathcal{L}_2$ dans laquelle X n'apparaît pas librement.

Quelques remarques avant de poursuivre : premièrement, insistons sur le fait que dans la plupart des travaux contemporains $\mathcal{Z}_2$ est traitée comme une théorie *sortée du premier ordre*. Si à l'inverse on interprète la quantification sur les variables d'ensembles $(\exists X, \forall X)$ au moyen de la sémantique classique, pleine, de la logique du second ordre [146], ce qui

145. Pour mémoire :

$$\begin{array}{l}\forall x(0 \neq s(x))\\ \forall x\forall y(s(x) = s(y) \rightarrow x = y)\\ \forall x(x \neq 0 \rightarrow \exists y(s(y) = x))\\ \forall x(x + 0 = x)\\ \forall x\forall y(x + s(y) = s(x + y))\\ \forall x(x.0 = 0)\\ \forall x\forall y(x.s(y) = x.y + x)\\ \forall x\forall y\Big(x < y \leftrightarrow (\exists z(x + s(z) = y))\Big)\end{array}$$

Notez que toutes les quantifications portent sur des variables du premier ordre (*i.e.* de la première sorte).

146. nous voulons parler ici de la sémantique standard ou full second order semantics, par opposition à la sémantique de Henkin

revient à exiger que les variables de second ordre puisse prendre pour valeur tous les sous-ensembles possibles du domaine sur lequel portent les variables individuelles de premier ordre, on obtient l'arithmétique de Peano au second ordre PA^2 qui est bien connue pour être catégorique.[147] Remarquez que les axiomes de compréhension deviennent aussitôt superflus puisque cette sémantique implique que toutes les sous-parties de l'univers de discours existent et peuvent être prises comme valeurs par les variables d'ensemble. Bien entendu, cette sémantique ne peut être munie d'un système de preuve (récursif) complet.

C'est précisément pourquoi dans le cadre d'une étude preuve-théorique, il est préférable de traiter $\mathcal{Z}_2$ comme une théorie du premier ordre sortée.[148] Dans ce cas, un modèle de $\mathcal{Z}_2$ est une structure $\mathfrak{M} := \langle M, \mathcal{S}(M), 0^{\mathfrak{M}}, +^{\mathfrak{M}}, .^{\mathfrak{M}}, s^{\mathfrak{M}}, <^{\mathfrak{M}} \rangle$[149] où M est un ensemble formant le domaine sur lequel les variables numériques (de la première sorte) $x, y, \ldots$ prennent leurs valeurs, tandis que $\emptyset \subsetneq \mathcal{S}(M) \subseteq \mathcal{P}(M)$ est un sous-ensemble, éventuellement strict mais toujours non vide, de l'ensemble des parties de M, sur lequel les variables d'ensembles (de la seconde sorte) $X, Y, \ldots$ prendront leurs valeurs. On retrouve alors de bonnes propriétés preuve-théoriques (complétude, compacité, *etc.*). La catégoricité est en revanche perdue.[150]

La formulation au premier ordre permet en particulier un contrôle fin des hypothèses concernant l'existence d'ensembles que nous sommes prêts à accepter à travers les axiomes de compréhension. Ainsi tout modèle $\mathfrak{M} := \langle M, \mathcal{S}(M), \ldots \rangle$ de $\mathcal{Z}_2$ contiendra *a minima* dans $\mathcal{S}(M)$, tous les ensembles définissables dans $\mathcal{L}_2$.[151] Mais on pourrait très bien relâcher cette contrainte en affaiblissant le schéma d'axiome de compréhension. On obtiendra alors un sous-fragment de $\mathcal{Z}_2$, c'est-à-dire une théorie exprimée dans $\mathcal{L}_2$ dont tous les axiomes sont des théorèmes de $\mathcal{Z}_2$. Par exemple, si on stipule que seules des formules arithmétiques, c'est-à-dire des formules dans lesquelles aucune quantification du second ordre n'apparaît, peuvent être employée dans le schéma d'axiomes de compréhension *iii*) ci-dessus, et qu'on formule l'induction au moyen de l'axiome du second ordre

147. PA^2 admet donc un unique modèle à isomorphisme près qui n'est autre que le modèle standard : $\mathbb{N} := \langle \omega, \mathcal{P}(\omega), +^{\mathbb{N}}, .^{\mathbb{N}}, s^{\mathbb{N}}, <^{\mathbb{N}} \rangle$

148. Ou, pour le dire de manière quasi équivalente, il est préférable de munir notre théorie d'une sémantique à la Henkin.

149. Notez que $\in$ est toujours traitée comme la « vraie » appartenance.

150. Le modèle standard ou attendu de $\mathcal{Z}_2$ est $\mathbb{N} := \langle \omega, \mathcal{P}(\omega), +^{\mathbb{N}}, .^{\mathbb{N}}, s^{\mathbb{N}}, <^{\mathbb{N}} \rangle$. Mais $\mathcal{Z}_2$ admet de nombreux modèles non standard. Un ω-modèle est un modèle dont l'univers sous-jacent (où les variables de la première sorte prennent leurs valeurs) est l'ensemble des entiers naturels. Pour (beaucoup) plus de détails sur les divers modèles de $\mathcal{Z}_2$, en tant que théorie sortée du premier ordre, voir Simpson (2009).

151. Ce qui dans bien des cas ne représente qu'une petite partie de $\mathcal{P}(M)$, puisqu'il n'y a qu'un nombre dénombrable d'ensembles définissables dans $\mathcal{L}_2$.

(la version *ii*) ci- dessus) on obtient le sous-système de $\mathcal{Z}_2$ dit de compréhension arithmétique et habituellement noté ACA_0.[152] ACA_0 contient donc comme axiome d'induction la formule suivante :

$$(0 \in X \wedge \forall n(n \in X \rightarrow s(n) \in X)) \rightarrow \forall n(n \in X)$$

et comme axiomes de compréhension toutes les clôtures universelles de formules suivantes :

$$\exists X \forall n(n \in X \leftrightarrow \varphi(n))$$

où $\varphi(n)$ est une formule quelconque de $\mathcal{L}_2$ dans laquelle X n'apparaît pas et qui ne contient aucune variable du second ordre quantifiée. Notez en revanche que $\varphi(n)$ peut contenir des variables du second ordre (autres que X) libres. Ces dernières sont en fait considérées et traitées comme des paramètres. Il y a bien sûr de nombreuses autres possibilités résultants en des hypothèses d'existence d'ensembles plus ou moins strictes ou plus moins larges, et permettant de définir d'autres sous-systèmes de $\mathcal{Z}_2$.[153]

Lorsqu'on définit un sous-système de $\mathcal{Z}_2$ la formulation de l'induction est cruciale. En effet, si l'on formule l'induction sous la forme d'un unique axiome employant une variable du second ordre (type *ii*) ci-dessus), alors la force de cet axiome va être directement tributaire des axiomes de compréhension que l'on a acceptés préalablement et qui seuls permettent de construire les ensembles que la variable du second ordre de

152. ACA pour « arithmetical comprehension axioms », le 0 en indice indiquant quant à lui que l'induction est également restreinte (voyez ci-desssous).

153. Les cinq principaux sous-systèmes de $\mathcal{Z}_2$ qui ont été isolés et étudiés dans la littérature de la théorie réductive de la preuve sont, par ordre croissant de force logique : $RCA_0, WKL_0, ACA_0, ATR_0$, et $\Pi^1_1 - CA_0$. Nous les donnons ici simplement à titre informatif. Nous ne faisons bien entendu qu'à peine effleurer la surface de ce domaine de recherche en mathématiques et en philosophie des mathématiques. Signalons toutefois encore le point suivant assez frappant : selon Simpson (2009, chapitre 1, p. 46-47), ces cinq principaux sous-fragments de $\mathcal{Z}_2$ qui ont été isolés pour des raisons avant tout « techniques » de reconstruction de portions de plus en plus inclusives des mathématiques ordinaires, correspondent à diverses positions bien connues et philosophiquement argumentées en matière de fondations des mathématiques. Simpson donne la typologie suivante :

RCA_0	constructivisme	Bishop
WKL_0	réductionnisme finitiste	Hilbert
ACA_0	prédicativisme	Weyl, Feferman
ATR_0	réductionnisme prédicatif	Friedman, Simpson
$\Pi^1_1 - CA_0$	imprédicativité	Ferferman *et al.*

l'axiome d'induction pourra prendre comme valeur. [154] Mais dans bien des cas on peut souhaiter faire varier l'induction de manière indépendante de la manière dont varie la compréhension, soit qu'on veuille restreindre l'induction à une classe d'ensembles plus réduite que celle que l'on accepte dans la compréhension, soit au contraire qu'on veuille pouvoir employer l'induction sur une classe d'ensembles plus large que celle que l'on pourra construire en employant un schéma de compréhension fixé par ailleurs. Dans de pareils cas, la formulation de l'induction au moyen d'un schéma d'axiomes (version ii') ci-dessus) sera préférable. On pourra en effet stipuler de manière indépendante la classe des formules pouvant être employées dans le schéma d'axiomes d'induction et celle des formules pouvant être employées dans le schéma d'axiomes de compréhension. [155] C'est en particulier le cas pour un sous-fragment de $\mathcal{Z}_2$ notablement important pour notre discussion.

Définition. ACA : Le sous-système de $\mathcal{Z}_2$ baptisé ACA est obtenu en prenant

— l'ensemble des axiomes de base i),

— le schéma ii') d'axiomes d'induction du second ordre complet, c'est-à-dire *toutes* les formules de la forme :

$$\Big(\varphi(0) \wedge \forall n(\varphi(n) \to \varphi(s(n)))\Big) \to \forall n\varphi(n)$$

où φ est une formule *quelconque* de $\mathcal{L}_2$, [156]

154. Ainsi, dans ACA_0, à travers la restriction du schéma de compréhension, on a également une restriction de l'induction aux seuls ensembles définissables par une formule arithmétique. On peut d'ailleurs donner une formulation équivalente à ACA_0 en gardant le même schéma de compréhension restreint aux formules arithmétiques et en remplaçant l'axiome unique d'induction

$$(0 \in X \wedge \forall n(n \in X \to s(n) \in X)) \to \forall n(n \in X)$$

par le schéma d'axiomes

$$\Big(\varphi(0) \wedge \forall n(\varphi(n) \to \varphi(s(n)))\Big) \to \forall n\varphi(n)$$

où φ est une formule *arithmétique* de $\mathcal{L}_2$, *i.e.* sans quantificateur du second ordre mais avec éventuellement des variables libres du second ordre.

155. À titre d'exemple, le sous-système RCA_0, si fondamental pour les mathématiques inversées, est constitué des axiomes de base i) d'un schéma d'axiomes d'induction ii') pour toute les formules Σ^0_1 et d'un schéma d'axiomes de compréhension iii) restreint aux formules Δ^0_1, qui est une classe strictement contenue dans Σ^0_1. Le système ACA dont Halbach rappelle qu'il est lié intimement à PA^{+ind}_{Tar} fournit un autre exemple de sous-système de $\mathcal{Z}_2$ où induction et compréhension sont limitées à des classes différentes de formules.

156. Autrement dit, on ne restreint pas ici l'induction aux seules formules arithmétiques. C'est ce qui

— et le schéma *iii*) d'axiomes de compréhension *restreint* aux formules *arithmétiques*.

Comme le rappelle Halbach (2014), le lien de tout ceci avec nos extensions aléthiques est établi par le résultat suivant.

Théorème 19. *Les systèmes PA^{+ind}_{Tar} et ACA sont preuve-théoriquement équivalents. Plus précisément,*

1. *d'une part, il est possible de définir dans ACA un prédicat de vérité satisfaisant PA^{+ind}_{Tar} ,*
2. *et d'autre part, il existe une interprétation relative de ACA dans PA^{+ind}_{Tar} qui préserve les expressions arithmétiques (à un renommage de variables près).*

De là, il suit que ACA et PA^{+ind}_{Tar} prouvent exactement les mêmes énoncés de $\mathcal{L}_{PA}$ (i.e. du premier ordre).

Halbach déclare que les résultats de ce théorème sont notoirement connus des praticiens de la théorie de la preuve, qu'ils appartiennent en quelque sorte au folklore du domaine. La définition de la vérité pour PA dans ACA est évidemment très proche d'une construction à la Tarski, quoique ACA soit plus faible qu'une logique du second ordre standard. Halbach mentionne Takeuti (1987) pour une construction entièrement développée d'une telle définition. Pour ce qui est de l'interprétation relative d'ACA dans PA^{+ind}_{Tar}, Halbach renvoie à Feferman (1991) [157] où une telle traduction est pour la première fois ébauchée par écrit. Halbach (2014, § 8.6) contient une démonstration assez détaillée de ces deux résultats. Le premier n'est peut-être guère étonnant dans un monde post-tarskien ; le second en revanche est plus surprenant. Sans entrer plus avant dans les détails techniques [158] de l'interprétation de ACA dans PA^{+ind}_{Tar}, donnons néanmoins quelques indication sur sa construction.

On définit une fonction de traduction $^* : \mathcal{L}_2 \longrightarrow \mathcal{L}_{PA} \cup \{Vr\}$ qui envoie injectivement toutes les variables du premier ordre de $\mathcal{L}_2$ sur un sous-ensemble strict des variables de $\mathcal{L}_{PA} \cup \{Vr\}$ et toutes les variables du second ordre sur un autre sous-ensemble (disjoint du précédent) des variables de $\mathcal{L}_{PA} \cup \{Vr\}$. [159] Les autres symboles du langage de l'arithmétique (*i.e.* $0, +, ., s, <$) sont laissés inchangés. À l'aide de cette traduction on peut alors

distingue ACA de ACA_0. ACA est un système strictement plus fort.

157. ou plus précisément à une pré-version manuscrite non publiée de cet article.

158. pour lesquels nous renvoyons à nouveau à Halbach (2014).

159. Disons par exemple qu'on envoie les variables numérique de $\mathcal{L}_2$ sur les variables d'indices pairs, et

exprimer la relation d'appartenance $t \in X$ dans $\mathcal{L}_{PA} \cup \{Vr\}$ au moyen du prédicat de vérité. Dans ACA, en effet, les axiomes de compréhension sont limitées aux formules $\varphi(n)$ arithmétiques (avec paramètres). Lorsque X prend pour valeur un sous-ensemble défini au moyen d'une formule φ, dire qu'un élément y appartient à X peut s'exprimer en disant que la formule φ est vraie de y. [160] L'idée centrale de la démonstration est donc de s'appuyer sur le prédicat de vérité pour réduire la relation d'appartenance par une quantification sur les formules.

Ce qui est particulièrement remarquable, c'est qu'avec cette traduction, les (traductions des) axiomes de compréhension de ACA deviennent des théorèmes de PA^{+ind}_{Tar}. Ainsi, implicitement (ou potentiellement modulo une traduction), PA^{+ind}_{Tar} contient donc une dose non négligeable de théorie des ensembles. Et ACA et PA^{+ind}_{Tar} sont deux théories intimement liées l'une à l'autre au point qu'on peut les considérer comme des variantes notationnelles l'une de l'autre. [161] On peut y reconstruire une part non négligeable des mathématiques, [162] soit de manière relativement directe dans ACA, soit de manière un peu plus contournée dans PA^{+ind}_{Tar} en s'appuyant sur la réduction à cette dernière de ACA.

---

les variables d'ensemble sur les variables d'indices impairs :

$$\begin{aligned} {}^* : \mathcal{L}_2 &\longrightarrow \mathcal{L}_{PA} \cup \{Vr\} \\ {x_n}^* &= x_{2n+2} \\ {X_n}^* &= x_{2n+1} \end{aligned}$$

160. Plus précisément (d'après Halbach (2014)), si $\dot{h}(x, y)$ représente une fonction (récursive) qui lorsqu'on lui donne en entrée le couple formé du code $\ulcorner \phi(x_0) \urcorner$ d'une formule à une variable libre et d'un entier n donne en sortie $\ulcorner \phi(\overline{n}) \urcorner$ le code la formule ϕ où toutes les occurences libres de x on été remplacées par le numéral $\overline{n}$, alors $t \in X$ peut se traduire par $(t \in X_n)^* = Vr(\dot{h}(x_{2n+1}, t^*))$ où t^* est la traduction du terme numérique t (c'est-à-dire t où on a renommé les variables selon *). Lorsque x_{2n+1} prend pour valeur le code d'une formule $\varphi(x)$ à une variable libre, $Vr(\dot{h}(x_{2n+1}, t^*))$ exprime simplement le fait que $\varphi(t^*)$ est vraie, autrement dit le fait que t^* satisfait φ. Il faut ensuite encore prendre garde à bien relativiser les quantifications de $\mathcal{L}_2$ portant sur des variables d'ensemble lorsqu'on les traduit dans $\mathcal{L}_{PA} \cup \{Vr\}$, ce qui se fait en précisant que ces quantifications portent sur des (variables désignant des) formules... Ainsi $(\forall X_n \phi)^* = \forall x_{2n+1}(Form(x_{2n+1}, \ulcorner x_0 \urcorner) \rightarrow \phi^*)$ où $Form(x, y)$ exprime le fait que x est le code d'une formule ayant y comme unique variable libre. Voyez Halbach (2014, p 95-98) pour plus de détails.

161. Nous reprenons cette formulation à Halbach :

> [...] PA^{+ind}_{Tar} et ACA prouvent les mêmes théorèmes arithmétiques. Ces deux théories peuvent même être considérées comme des variantes notationnelles l'une de l'autre. PA^{+ind}_{Tar} est donc, *grosso modo*, [une manière de] parler des ensembles arithmétiques.
>
> Halbach (2001b, p. 187)

162. Puisque $ACA_0 \subset ACA$, au moins autant si ce n'est plus que la part prédicative des mathématiques selon Simpson (2009) (voyez le tableau de la note 153).

Une fois de plus, les résultats techniques ne semblent guère favorables à la stratégie avancée par Field (1999). Comme le rappelle Halbach, ACA est une extension non négligeable de PA. Elle est en particulier beaucoup plus forte que la théorie obtenue par le simple ajout de G ou de $Con(PA)$.

> [...] ACA permet la dérivation de nombreux résultats qui ne sont pas prouvables dans PA. [$Con(PA)$] n'est qu'une minuscule conséquence de ACA. L'énoncé de la consistance ajouté à PA n'engendre pas un système qui soit siginificativement plus fort que PA soi-même. Ces deux théories possèdent le même ordinal preuve-théorique [163], et l'énoncé de la consistance est également inutile pour établir des principes combinatoires qui ne sont pas déjà prouvables dans PA. Ceci contraste fortement avec ACA elle-même : ACA décide divers problèmes combinatoires qui ne sont pas prouvables dans PA ; elle possède un ordinal preuve-théorique bien plus élevé que PA ; elle prouve la consistance de certaines progressions de théories fondées sur le principe de réflexion uniforme —et [PA^{+ind}_{Tar}] peut en faire tout autant.
>
> (Halbach, 2001b, p. 187)

Et pour cause, au vu du résultat précédent, ce qui vaut pour ACA vaut aussi bien pour PA^{+ind}_{Tar}. Ainsi quand Field semble suggérer que le prédicat de vérité tel qu'il est axiomatisé par PA^{+ind}_{Tar} ne sert qu'à accroître le pouvoir expressif de notre langage pour nous permettre de renforcer l'induction et formuler par là une théorie arithmétique plus forte, il est tentant de lui rétorquer qu'il est pour le moins étonnant qu'une notion purement expressive ou non substantielle contienne *in fine* une telle mesure de théorie des ensembles. [164]

En outre, on peut affiner ce contre-argument car l'examen des résultats de théorie réductive de la preuve nous renseigne également sur le rôle de certains axiomes. Nous venons en effet de voir que ACA était une extension de PA preuve-théoriquement équivalente à PA^{+ind}_{Tar}. Elle est donc bien évidemment non conservative (syntaxiquement et modèle-théoriquement) sur PA. Néanmoins, si on restreint l'induction tout en gardant les axiomes de compréhension arithmétiques, par exemple en remplaçant dans ACA le schéma d'induction du second ordre complet :

163. En anglais : « proof-theoretic ordinal », au sens de la théorie de la preuve qui attribue des ordinaux aux systèmes d'axiomes en fonction de leur force logique.
164. C'est d'ailleurs à peu près ce que répond Shapiro (2003, p. 127).

$$(\varphi(0) \wedge \forall n(\varphi(n) \rightarrow \varphi(s(n)))) \rightarrow \forall n\varphi(n),$$
où φ est une formule *quelconque* de $\mathcal{L}_2$

par le schéma d'induction restreint aux seules formules arithmétiques :

$$(\varphi(0) \wedge \forall n(\varphi(n) \rightarrow \varphi(s(n)))) \rightarrow \forall n\varphi(n),$$
où φ est une formule de $\mathcal{L}_2$ *ne contenant aucune quantification du second ordre*

on obtient le système ACA_0 qui est bien connu pour être conservatif sur PA.[165] Dès lors, on pourrait tout aussi bien proposer une argumentation *à la* Field pour « déflater » une bonne dose de théories des ensembles. Voici à quoi une telle argumentation pourrait ressembler :

> Partant de la théorie PA et du langage $\mathcal{L}_{PA}$, on peut enrichir notre langage et étendre notre théorie pour obtenir le système ACA exprimé dans $\mathcal{L}_{PA} \subsetneq \mathcal{L}_2$. ACA est une extension non conservative de PA. Mais remarquez que tant que vous n'élargissez pas votre schéma d'induction au(x) (formules du) langage enrichi, vous obtenez ACA_0,[166] qui est une extension conservative de PA. Les axiomes que vous avez ajoutés pour passer de PA à ACA_0, ce qu'on pourrait appeler « les axiomes essentiels de l'appartenance » puisqu'ils ne dépendent que des notions nouvellement introduites[167] et non pas de l'induction, sont parfaitement acceptables d'un point de vue déflationniste. Les propriétés ou les notions qu'ils axiomatisent sont « non substantielles » et « non explicatives », ce dont témoigne la conservativité de la théorie avant extension de l'induction. Bien sûr, en enrichissant votre langage, vous avez

165. La preuve de la conservativité de ACA_0 sur PA consiste d'ailleurs à montrer que tout modèle de PA peut être étendu en un modèle de ACA_0 (en prenant comme domaine d'interprétation des variables d'ensembles $\mathcal{S}(M) = Def(M)$ l'ensemble des sous-parties de M définissables (avec paramètres) dans M). Ceci établit donc en fait la conservativité modèle-théorique de ACA_0 sur PA, ce qui, nous l'avons vu, est plus fort que la conservativité syntaxique. Un corollaire immédiat de ce résultat et du théorème de Lachlan (17) est que ACA_0 (contrairement à ACA) ne peut contenir de définition d'un prédicat de vérité pour $\mathcal{L}_{PA}$ satisfaisant les axiomes de PA_{Tar} !

166. Et même un sous-système un peu faible si l'on restreint vraiment aux seules formules de $\mathcal{L}_{PA}$ plutôt qu'aux formules arithmétiques de $\mathcal{L}_2$.

167. en l'occurrence ici, une nouvelle sorte de variables et un nouveau symbole de relation $\in$.

> augmenté le pouvoir expressif de ce dernier et vous pouvez donc à présent renforcer l'induction en obtenant de nouvelles instances de votre schéma dans lesquelles le vocabulaire du langage enrichi apparaît. Dès lors, vous êtes en position de pouvoir « formuler rigoureusement une théorie arithmétique plus forte que celle que nous pouvions rigoureusement formuler auparavant. Il n'y a rien ici de véritablement spécial à propos de la ~~vérité~~ [théorie des ensembles] : en employant n'importe quelle autre notion non exprimable dans le langage d'origine, nous pouvons obtenir de nouvelles instances du schéma d'induction et bien souvent ces dernières déboucheront sur des extensions non conservatives.

À ce compte, on a donc « déflaté » une part non négligeable de théorie des ensembles et donc de mathématiques. Une telle analyse semble extrêmement contre-intuitive. Pour le dire franchement, elle n'est pas plausible une seule seconde à nos yeux. Elle apparaît en tout cas en totale contradiction avec la pratique des mathématiciens, en particulier ceux qui, justement, tentent de préciser les hypothèses substantielles ou les engagement ontologiques que les diverses hiérarchie de systèmes d'axiomes recèlent. Du point de vue de la théorie réductive de la preuve, les hypothèses ensemblistes formant les axiomes de compréhension de ACA sont « par excellence » des hypothèses censées être substantielles et pouvant jouer un rôle explicatif. Étant donné la proximité d'ACA et de PA^{+ind}_{Tar}, on ne voit pas pourquoi il en serait autrement en ce qui concerne le prédicat de vérité axiomatisé par les clauses tarskiennes. [168]

Pour terminer cette excursion à travers les sous-systèmes de $\mathcal{Z}_2$, il y a encore un point qui nous semble pertinent pour la discussion de la stratégie fieldienne. Nous avons vu que PA^{+ind}_{Tar} était preuve-théoriquement équivalente à ACA, et que lorsqu'on affaiblissait suffisamment l'induction de ACA, on pouvait obtenir un sous-système de $\mathcal{Z}_2$ conservatif sur PA (typiquement : ACA_0). Si au lieu d'affaiblir l'induction d'ACA, on restreint au contraire les axiomes de compréhension, on peut construire un autre sous-fragment de $\mathcal{Z}_2$ remarquable pour l'analyse des extensions aléthiques de PA.

Définition. $ACA_{\restriction \mathcal{L}_{PA}}$: Le sous-système de $\mathcal{Z}_2$ composé des axiomes de base, du schéma

168. Sauf peut-être à considérer que les principes méthodologiques employés en théorie de la preuve et qui, pour une bonne part, sont hérités du programme de Hilbert, ne s'appliquent pas lorsqu'il s'agit d'évaluer les thèses déflationnistes sur la vérité. Peut-être. Mais on se demande bien ce qui justifie une tel traitement de faveur... (en annexe 4.4.1 nous avons tenté d'imaginer ce qu'aurait pu donner un dialogue historique (bien évidemment fictif et parodique) entre un Cantor « déflationniste » en matière de théorie des ensembles et un Hilbert tentant de mener à bien son Programme).

d'induction du second ordre complet (*i.e.* pour toutes les formules de $\mathcal{L}_2$, sans restriction) et du schéma d'axiomes appliqué uniquement aux formules du langage du premier ordre $\mathcal{L}_{PA}$ (*i.e.* aux formules ne contenant aucune variable d'ensembles, libre ou quantifiée) est appelé $ACA_{\restriction \mathcal{L}_{PA}}$.[169]

Bien évidemment, $ACA_{\restriction \mathcal{L}_{PA}}$ est inclus dans ACA et c'est un sous-système de $\mathcal{Z}_2$ strictement plus faible. $ACA_{\restriction \mathcal{L}_{PA}}$ est lié aux versions formalisées de la théorie *décitationnelle* de la vérité d'une manière semblable à celle dont ACA est lié à l'axiomatisation tarskienne PA^{+ind}_{Tar}. Plus précisément, le sous-fragment $ACA_{\restriction \mathcal{L}_{PA}}$ de $\mathcal{Z}_2$ est interprétable non pas dans PA_d mais dans une version légèrement renforcée de celui-ci qu'on peut appeler théorie de la décitation uniforme :

Définition. Extension aléthique décitationnelle uniforme : Soit T une théorie exprimée dans $\mathcal{L}_T$ et capable de représenter sa propre syntaxe au moyen d'un codage gödelien. On appelle **T**-équivalences uniformes l'ensembles des biconditionnels suivants :

$$\forall x_1, \ldots, \forall x_n (Vr(\ulcorner \varphi(\mathring{x_1}, \ldots, \mathring{x_n}) \urcorner) \leftrightarrow \varphi(x_1, \ldots, x_n))$$

où $\varphi(x_1, \ldots, x_n)$ est une formule de $\mathcal{L}_T$.[170]

Si on part d'une théorie de base arithmétique contenant un schéma d'axiomes d'induction (disons par exemple PA), on obtiendra deux extension aléthiques décitationnelles uniformes selon qu'on étendra ou non l'induction au vocabulaire de $\mathcal{L}_T \cup \{Vr\}$:

1. $T_{d.u.} :=$

   $$T \cup \{\forall x_1, \ldots, \forall x_n \Big(Vr(\ulcorner \varphi(\mathring{x_1}, \ldots, \mathring{x_n}) \urcorner) \leftrightarrow \varphi(x_1, \ldots, x_n) \Big) \mid \varphi(\vec{x_i}) \in Form(\mathcal{L}_T)\}$$

   et le schéma d'induction est traité comme *liste*.

2. $T^{+ind}_{d.u.} :=$

   $$T \cup \{\forall x_1, \ldots, \forall x_n \Big(Vr(\ulcorner \varphi(\mathring{x_1}, \ldots, \mathring{x_n}) \urcorner) \leftrightarrow \varphi(x_1, \ldots, x_n) \Big) \mid \varphi(\vec{x_i}) \in Form(\mathcal{L}_T)\}$$

   et le schéma d'induction est traité comme *règle*.

Avec cette notation précisée, on a le résultat suivant dû à Halbach (2014, 1999a) :

169. Halbach (2014, 1999a) le baptise ACA_{PF} pour axiomes de compréhension arithmétique sans paramètre (PF pour *Parameter Free*).

170. Il s'agit donc bien là encore d'une axiomatisation pour un prédicat de vérité *typé* : ce dernier ne s'applique qu'à des formules ne contenant pas déjà elle même le prédicat « vrai ».

Théorème 20. *$ACA_{\restriction \mathcal{L}_{PA}}$ est relativement interprétable dans $PA^{+ind}_{d.u.}$, l'extension aléthique décitationnelle uniforme de PA avec induction étendue.*

En outre, comme $PA^{+ind}_{d.u.}$ est conservative sur PA, on peut en déduire également que $ACA_{\restriction \mathcal{L}_{PA}}$ est une extension conservative de PA

À la suite de ce théorème, il vaut la peine de remarquer que si l'extension aléthique $PA^{+ind}_{d.u.}$ est suffisamment forte pour qu'on puisse y interpréter $ACA_{\restriction \mathcal{L}_{PA}}$, elle est en revanche trop faible pour qu'on puisse réaliser en son sein une interprétation relative d'ACA. En fait, les axiomes compositionnels $\{Tar\}$ compris dans l'extension PA^{+ind}_{Tar} sont indispensables pour pouvoir mener à bien la traduction des axiomes de compréhension d'ACA qui contiennent des formules de $\mathcal{L}_2$ arithmétiques à paramètres (*i.e.* des formules du second ordre sans variables du second ordre quantfiées), et prouver (les traductions de) ces axiomes à l'intérieur de PA^{+ind}_{Tar}.[171] En d'autres termes, alors même qu'ils ont exactement les mêmes axiomes d'induction, ce qui distingue $PA^{+ind}_{d.u.}$ et PA^{+ind}_{Tar}, c'est que la présence des clauses compositionnelles pour le prédicat de vérité permet d'exprimer (et de dériver) des hypothèses concernant l'existence de certains ensembles[172] beaucoup plus forte que celles exprimables et dérivables dans une extension par un prédicat strictement (uniformément) décitationnel. Et la différence n'est pas minime : ACA contient des hypothèses ensemblistes bien plus fortes que celles de $ACA_{\restriction \mathcal{L}_{PA}}$.

Voilà, semble-t-il, une raison supplémentaire de douter que les axiomes $\{Tar\}$, les « axiomes essentiels » de la vérité, sont, comme paraît le suggérer Field (1999), sans contenu substantiel et sans effet explicatif, et qu'ils jouent simplement un rôle d' « auxiliaire expressif » dans l'apparition des phénomènes de non conservativité. Ce sont bien ces axiomes qui implicitement nous engagent envers des hypothèses ensemblistes plus fortes que celles auxquelles nous engagent les théories décitationnelles.

4.2.2.4 Conclusion sur Field (1999)

Pour conclure, la réaction de Field (1999) au dilemme face auquel Shapiro et Ketland entendaient placer le déflationniste, semblait particulièrement ingénieuse. En distinguant les « axiomes essentiels » de la vérité et les axiomes d'induction élargie, Field paraissait, presque miraculeusement, pouvoir rendre compatibles les deux contraintes de

171. Voyez Halbach (2014, p. 97) pour plus de détails.
172. ceux qui sont définissables par des formules arithmétiques à paramètre de $\mathcal{L}_2$.

conservativité et de réflexivité. Le point clef de l'argumentation de Field était évidemment la conservativité des axiomes compositionnels PA_{Tar} sur PA. Malheureusement, l'astucieuse construction de Field ne nous semble pas résister à une analyse plus poussée. Nous avons en effet exposé plusieurs contre-arguments qui sont à nos yeux assez dévastateurs pour l'argumentation de Field. Rappelons en les principaux aspects.

Lorsqu'en s'appuyant sur la conservativité (syntaxique) de PA_{Tar} sur PA, Field prétend que les axiomes aléthiques ne jouent pas de rôle explicatif dans la dérivation du principe de réflexion et que la notion de vérité ainsi formalisée n'est qu'un simple outil expressif permettant de renforcer l'induction et de formuler une théorie arithmétique plus forte, on peut lui rétorquer que les axiomes compositionnels jouent bel et bien un rôle crucial dans les arguments sémantiques : renforcer l'induction au moyen d'un prédicat de vérité purement décitationnel ne suffit pas à augmenter nos capacités de preuves, ce dont témoigne la conservativité de PA_d sur PA. Le point crucial est que l'ensemble des énoncés vrais est clos par application des règles de déduction de PA ; et pour établir ceci, il faut à la fois une induction élargie *et* les clauses tarskiennes pour le prédicat « vrai ».

À cela s'ajoute le fait que s'il est exact que PA_{Tar} est syntaxiquement conservative sur PA, cela ne signifie pas que l'ajout des clauses $\{Tar\}$ est sans impact sur la structure possible du monde (ou sur la classe des modèles satisfaisant la théorie étendue). Bien au contraire, en employant des outils plus sophistiqués de théorie des modèles, on peut constater que PA_{Tar} n'est pas modèle-théoriquement conservative sur PA, et ce indépendamment de tout élargissement de l'induction. Nous avons d'ailleurs suggéré que si PA_{Tar} ne permet pas de dériver un nouveau théorème de $\mathcal{L}_{PA}$ laissé indécidé par PA, ce n'est pas parce que la notion de vérité est métaphysiquement neutre, non substantielle ou non explicative, c'est plutôt parce qu'ici pouvoir expressif et pouvoir explicatif sont inséparables, au sens où les conséquences substantielles et bien réelles de $\{Tar\}$ sur la structure des modèles possibles ne sont pas exprimables dans le langage d'origine.

Enfin, le lien fait avec des résultats classiques de théorie réductive de la preuve a permis de montrer que PA^{+ind}_{Tar} était une extension forte de PA. Elle est en effet preuve-théoriquement équivalente au sous-fragment ACA de l'arithmétique du second ordre. PA^{+ind}_{Tar} contient donc implicitement une bonne dose de théorie des ensembles. Là encore, les axiomes compositionnels sont centraux pour traduire la relation ensembliste d'appartenance et permettre d'exprimer l'équivalent d'axiomes de compréhension forts, au point d'obtenir une « variante notationnelle » d'ACA.

Au total, la tentative de concilier les deux branches du dilemme (*i.e.* la contrainte de conservativité et celle de réflexivité) en distinguant les axiomes essentiels de la vérité nous semble donc sans espoir.

Dans la section 4.1, nous avons déjà discuté la contrainte de conservativité et défendu l'idée qu'au vu des thèses qu'il défend concernant le prédicat de vérité, une telle contrainte s'imposait bel et bien au déflationniste. Pour sortir du piège tendu par Shapiro et Ketland, il reste toutefois une possibilité pour le déflationniste : remettre en cause l'autre prémisse du dilemme, à savoir la contrainte de réflexivité. Sommes-nous en effet certains qu'une théorie adéquate de la vérité doit permettre de justifier le schéma de réflexion (Ref) et, de ce fait, fournir une démonstration sémantique de G et $Con(PA)$? Peut-être y a-t-il d'autres moyens de parvenir à ces conclusions. C'est en tout cas ce qu'affirme Tennant (2002). Et c'est à présent vers une analyse de cette réponse fournie au nom du déflationnisme que nous nous tournons.

4.3 La contrainte de réflexion revisitée et rediscutée

Souvenons-nous que pour Ketland (1999) et Shapiro (1998b), la contrainte de réflexivité portant sur la vérité revenait à exiger que, pour une théorie de base T, il soit possible de fournir dans une extension aléthique adéquate de cette théorie une *justification* de l'énoncé

(Ref) : « tous les théorèmes de T sont vrais ».

Typiquement, nous l'avons vu, cette justification prenait la forme d'une démonstration de l'énoncé (Ref) menée au moyen d'un argument inductif portant sur la longueur des preuves dans T. Selon Shapiro et Ketland, c'est cette capacité à *prouver* le principe de réflexion qui constitue la contrainte de réflexivité et forme la seconde prémisse de leur argument anti-déflationniste.

Il faut bien admettre que cette contrainte d'adéquation a souvent été considérée comme allant de soi par les philosophes ou logiciens s'intéressant à la vérité, en particulier par ceux qui souscrivent au projet d'en élaborer une théorie formalisée. Dès les années trente, Tarski (1935) considérait comme un succès majeur de sa définition (dans une métathéorie) de la vérité qu'elle permette de prouver la cohérence de la théorie objet, tandis qu'à l'inverse il rejetait une axiomatisation du prédicat de vérité limitée aux seules **T**-équivalences , au motif qu'

> une théorie de la vérité fondée sur elles serait un système hautement incomplet, auquel manquerait les plus importants et les plus fructueux théorèmes généraux. (Tarski, 1935, p. 257)

Pour prendre un exemple plus contemporain : de nos jours encore, dans un article de synthèse où il dresse la liste des propriétés que les théories formelles de la vérité sont à première vue censées posséder, Leitgeb (2007) [173] place la réflexivité en seconde position parmi huit caractéristiques souhaitables :

> (b) Si une théorie de la vérité est ajoutée à des théories mathématiques ou empiriques, il devrait être possible de *prouver* que ces dernières sont vraies.
>
> *C'est un point qui ne prête pas à controverse.* [...] En fait, il devrait non seulement être possible de prouver que chaque théorème de T pris isolément est vrai, mais *l'énoncé général affirmant que tout théorème de* T *est vrai, i.e.* 'pour tout x, si x est prouvable dans T alors $Vr(x)$', devrait être dérivable dans une théorie de la vérité adéquate. [...] En résumé : une théorie de la vérité devrait être élaborée de manière à ce que si la vérité doit être expliquée pour le langage d'une certaine théorie T, alors ajouter à T une telle théorie de la vérité devrait nous permettre de prouver la vérité (des membres) de T [...] (Leitgeb, 2007, p. 277-278, nous soulignons)

Le caractère apparemment banal et bien enraciné de la contrainte de réflexivité explique peut-être pourquoi ni Shapiro (1998b) ni Ketland (1999) ne se sont attardés à expliquer en détails en quoi une telle contrainte s'imposait à une théorie satisfaisante de la vérité, alors même qu'ils ont discuté abondamment la contrainte de conservativité. C'est peut-être également la raison de la réaction quelque peu laconique de Shapiro (2003) face aux défenses du déflationnisme qui proposent d'abandonner cette contrainte :

> [...] Tennant (2002) dépeint le déflationniste comme objectant ouvertement à des généralisation du type $[(Ref)]$, simplement parce qu'elles ne découlent pas des instances du schéma de vérité [N.D.T. *i.e.* des **T**-équivalences]. Pour la même raison le déflationniste devrait vraisemblablement objecter également à $[(InfVr)]$, l'affirmation que les règles d'inférences préservent la vérité. Tennant se donne beaucoup de mal pour montrer comment le déflationniste peut néanmoins obtenir l'effet de ces généralisations logiques,

173. Selon se propres termes, Leitgeb entend exposer dans cet article, « ce à quoi une théorie de la vérité *devrait* ressembler, du moins à première vue » (*cf.* Leitgeb (2007, abstract p. 276)).

> en acceptant certaines règles d'inférence [...]. Toutefois, j'aurais tendance à considérer comme un *reductio* qu'une explication de la vérité donnée ne nous permette pas d'établir que les règles d'inférence préservent la vérité [...] Mais peut-être sommes nous parvenus à un conflit d'intuitions. (Shapiro, 2003, p. 116)

Puis quelques pages plus loin :

> Selon [Azzouni (1999)] la déflationniste au premier ordre ne se donne pas assez de ressources pour établir que les théorèmes de la théorie de base sont vrais. Tout ce qu'elle se donne concernant les entiers naturels, combiné avec tout ce qu'elle se donne concernant la vérité, est insuffisant pour prouver que chaque théorème de $[PA]$ est vrai. Comme ci-dessus, j'aurais tendance à considérer ceci comme un *reductio ad absurdum* pour la déflationniste au premier ordre, mais je réalise que ce qui semble un modus tollens pour une personne semble un modus ponens pour une autre. (Shapiro, 2003, p. 125)

Il est clair que pour Shapiro, renoncer à la contrainte de réflexivité pour une théorie de la vérité adéquate est une option qui ne semble guère envisageable. Sans doute s'agit-il pour lui d'une contrainte d'adéquation « intuitivement incontestable » et s'imposant à toute tentative de formalisation de la vérité. Peut-être aussi qu'à ce titre, il serait à la fois vain et superflu de tenter de fournir une justification de cette contrainte, qui s'appuierait sur des principes plus fondamentaux ou plus évidents. Peut-être que cela fait partie de notre compréhension du concept de vérité que d'accepter comme indubitable que lorsqu'on ajoute une théorie de la vérité à une théorie de base il devrait être possible de prouver que cette dernière est vraie. Quoi qu'il en soit, qu'on partage ou non les convictions profondes de Shapiro, force est de constater qu'une fois de plus, comme Shapiro lui-même en convient, les intuitions concernant les propriétés fondamentales de la vérité ou les principes premiers censés en guider la formalisation sont loin d'être unanimement partagées. C'est en effet précisément à cette contrainte de réflexivité que Tennant (2002) propose de renoncer au nom du déflationnisme. Ketland a pour sa part vivement réagi à la réponse proposée par Tennant et il s'en est suivi un débat houleux entre ces deux auteurs tout d'abord sur le forum de discussion en ligne *Fondations of Mathematics (F.O.M)*, puis dans une série d'articles parus dans *Mind* Ketland (2005, 2010) et Tennant (2010, 2005). Nous allons donc à présent analyser en détails cette discussion. À nos yeux, la question fondamentale au coeur de cette controverse porte

en fait sur la justification du principe, ou plutôt des principes, de réflexion. Avant de l'examiner plus avant, il nous faut faire quelques rappels sur les diverses formes que ces principes peuvent prendre.

4.3.1 Schémas de réflexion

Dans le sillage de la parution des théorèmes d'incomplétude de Gödel, les principes ou schémas [174] de réflexion ont bien vite retenu l'attention et suscité l'intérêt des logiciens et mathématiciens. D'une part, les explications sémantiques du type de celles évoquées par Shapiro et Ketland établissant la vérité d'énoncés indécidables se sont rapidement répandues [175]. D'autre part, dans la mesure où l'ajout d'un schéma de réflexion à une théorie de base débouche la plupart du temps sur une extension non conservative, ce type de schéma est aussi apparu comme une manière naturelle de renforcer nos théories et de surmonter, au moins partiellement, les limites posées par l'incomplétude. [176] Les schémas de réflexion sont en effet généralement vus comme une manière d'exprimer notre confiance dans la fiabilité de nos méthodes de preuves. Si l'on accepte une théorie de base T, il peut donc sembler raisonnable (voire inévitable) d'accepter aussi non seulement tous les théorèmes de T mais également le ou les énoncé(s) qui en exprime(nt) la correction. Au moyen d'un prédicat de vérité pour le langage de T, on peut exprimer la fiabilité ou la correction [177] de T par l'énoncé (Ref) : « tous les théorèmes de T sont vrais ». Cet énoncé (Ref) est traditionnellement intitulé schéma de réflexion *global* pour T dans la

174. Suivant en cela la pratique de la littérature logique sur ce sujet, nous ne faisons pas de distinction entre les « principes » et les « schémas » de réflexion et nous employons les deux terminologies de manière indifférenciée. Peut-être serait-il préférable de réserver de l'expression *schéma* de réflexion à ce qu'on nomme généralement principe de réflexion local et principe de réflexion uniforme puisqu'il s'agit là effectivement de schémas d'axiomes (voyez ci-dessous), tandis que le principe global de réflexion est lui donné sous la forme d'un axiome unique (exprimé dans un langage enrichi). Dans la mesure où aucune confusion ne s'en suit nous n'avons pas pris cette précaution.

175. Nous avons déjà dit qu'à notre connaissance la première démonstration détaillée d'un énoncé de Gödel G_T pour une théorie arithmétique de base T, formalisée dans une extension aléthique de T et employant explicitement un schéma (global) de réflexion était due à Mostowski (1952). Signalons toutefois ici que ce type de démonstration était connu bien avant dès les années 1930. Dans le Postscript de son fameux article sur la concept de vérité paru en 1936, Tarski donne déjà l'ébauche d'une telle démonstration (voyez Tarski (1935, p. 275-276)).

176. Les premières tentatives systématiques en ce sens sont dues à Rosser (1937) et Turing (1939). Les idées avancées par Turing seront ensuite reprises et développées par Kreisel et Feferman dans les années 1960 pour aboutir à d'importants résultats en théorie de la preuve (voir Franzén (2004) pour une synthèse récente sur ce sujet).

177. Nous employons les termes « fiabilité » et « correction » à propos d'une théorie de façon interchangeable. Les deux expressions peuvent être vues comme des traductions du termes anglais « *soundness* » que l'on retrouve dans la littérature, majoritairement de langue anglaise, sur les schémas de réflexion.

littérature logico-mathématique.

Selon Halbach (2014, p. 322) :

> Le principe de réflexion global apparaît comme l'expression complète de l'affirmation de la correction de $[T]$, dans la mesure où il exprime que tous les théorèmes de $[T]$ sont vrais.

Sous cette forme, l'expression de la correction de T est donnée par un énoncé *unique*. Toutefois, recourir à (Ref) nécessite de se placer dans un langage plus riche que le langage de la théorie de base $\mathcal{L}_T$ puisque la notion de vérité n'est le plus souvent pas disponible, ni définissable dans $\mathcal{L}_T$. Cela requiert donc de s'appuyer sur des ressources lexicales et conceptuelles qui ne sont généralement pas déjà données dans T et son langage. Pour contourner ce problème les logiciens ont d'ordinaire eu recours à des schémas d'axiomes formulés dans le langage de base $\mathcal{L}_T$. La réflexion sur T prend alors la forme d'un ensemble *infini* d'énoncés. En outre, certains choix doivent alors être faits sur la manière dont ces schémas d'axiomes sont formulés et ces choix peuvent avoir un impact sur la force logique de l'extension obtenue. Depuis l'article classique de Feferman (1962), on distingue les schémas de réflexion *locaux* et *uniformes*, auxquels la plupart des autres formulations peuvent généralement se réduire [178].

Ainsi, soit T une théorie contenant assez d'arithmétique pour représenter sa propre syntaxe, par ordre *décroissant* de force logique nous avons les principes de réflexion suivants qui peuvent être ajoutés à T :

- Principe global de réflexion, exprimé par un énoncé unique dans une extension aléthique de T et noté $Ref(T)$

  $Ref(T)$ $\quad \forall x(Thm_T(x) \rightarrow Vr(x))$

- Principe de réflexion uniforme, exprimé dans le langage de T sous la forme d'un schéma d'axiomes, et noté RFN(T) :

  RFN(T) $\quad \forall x Thm_T(\ulcorner \varphi(\dot{x}) \urcorner) \rightarrow \forall x \varphi(x)$, où $\varphi(x)$ est un énoncé de $\mathcal{L}_T$ à une variable libre et où $\ulcorner \varphi(\dot{x}) \urcorner$ est un terme canonique définissable dénotant une fonction qui à tout entier n associe le code de $\ulcorner \varphi(\overline{n}) \urcorner$ obtenu en substituant le numéral $\overline{n}$ à la variable x dans $\varphi(x)$ [179].

178. Sous réserve que la théorie de base T à laquelle on ajoute ces schémas ne soit pas trop faible. Pour un examen plus poussé des diverses formulations des schémas de réflexion et des équivalences éventuelles entre ces diverses formulations voir Feferman (1962) ou plus récemment Beklemishev (2005).

179. Pour plus de détails sur cette convention d'écriture pointée « $\dot{x}$ » due à Feferman, voir Feferman (1960).

- Principe de réflexion local, exprimé dans le langage de T sous la forme d'un schéma d'axiomes, et noté Rfn(T) :

  Rfn(T) $\quad Thm_T(\ulcorner\varphi\urcorner) \to \varphi$, où φ est un énoncé de $\mathcal{L}_T$

Lorsqu'ils prennent forme de schémas d'axiomes exprimés dans le langage de base, Feferman qualifie les principes de réflexion comme étant

> une procédure pour ajouter à n'importe quel ensemble d'axiomes A, certains nouveaux axiomes dont la validité découle de la validité des axiomes A et qui expriment formellement, à l'intérieur du langage de A, des conséquences évidentes de l'hypothèse que tous les théorèmes de A sont valides. Feferman (1962, p. 274)

Sans entrer plus que nécessaire dans les détails techniques d'une analyse de la force logique des divers principes de réflexion, nous nous limiterons à rappeler les principaux résultats reliant les schémas ci-dessus. Pour simplifier la discussion, nous nous bornerons au cas où la théorie de base est l'arithmétique de Peano, *i.e.* $T = PA$ [180].

Tout d'abord, il est clair que le schéma de réflexion local est impliqué par le schéma de réflexion uniforme :

$$PA + \text{RFN(PA)} \vdash \text{Rfn(PA)}$$

Pour le voir, il suffit de remarquer que chaque instance de Rfn(PA) est aussi une instance de RFN(PA).

À l'inverse, le schéma de réflexion uniforme ne découle pas du schéma de réflexion local :

$$PA + \text{Rfn(PA)} \nvdash \text{RFN(PA)}$$

180. De nombreux raffinements et variations sont possibles : on peut considérer une théorie arithmétique de base T plus faible que PA, dès lors qu'elle peut tout de même représenter sa propre syntaxe et définir un prédicat de prouvabilité $Thm_T(x)$. L'arithmétique de Robinson $\mathcal{Q}$, ou bien l'arithmétique élémentaire (EA) sont des exemples possibles. On peut également prendre une théorie syntaxique d'arrière-plan différente de celle pour laquelle la prouvabilité est définie et examiner la force d'extensions du type $T_0 + Rfn(T_1)$ où $T_0 \neq T_1$. Enfin, on peut aussi limiter la classe Γ des énoncés de $\mathcal{L}_T$ sur laquelle des instances du schéma d'axiomes sont construites pour obtenir des extensions du type $T + \text{RFN}_\Gamma\text{(T)} := T \cup \{\forall x(Thm_T(\ulcorner\varphi(\dot{x})\urcorner) \to \varphi(x)) \mid \varphi(x) \in \Gamma\}$, où $\varphi(x)$ est un énoncé de $\mathcal{L}_T \cap \Gamma$ (typiquement on limite le schéma à une classe d'énoncé Σ_n ou Π_n d'énoncé de $\mathcal{L}_T$) . Nous laissons tous ces aspects techniques de côté. Les résultats que nous donnons dans ce qui suit sont tirés de Beklemishev (2005) et de Halbach (2014, chapitre 22.1).

Pour une démonstration de ce résultat classique et bien connu, nous renvoyons à Beklemishev (2005, p. 214) [181]. Voilà donc pour les rapports entre les deux formes principales de réflexion exprimées sans utiliser le prédicat de vérité : la réflexion uniforme (RFN) est en général strictement plus forte que la réflexion locale (Rfn).

Quels liens entre ces schémas et le principe *global* de réflexion formulé dans une extension aléthique ? Il est facile de voir que *modulo* la collection des **T**-équivalences , $PA + Ref(PA)$ prouve toutes les instances de Rfn(PA) :

$$PA_d + Ref(PA) \vdash \text{Rfn(PA)}$$

Il y a bien entendu une démonstration directe quasi triviale [182] de ceci :

Soit $\varphi \in \mathcal{L}_{PA}$, avec $PA_d = PA \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{PA}\}$, on a :

$$PA_d + \forall x(Thm_{PA}(x) \rightarrow Vr(x)) \vdash Thm_{PA}(\ulcorner\varphi\urcorner) \rightarrow Vr(\ulcorner\varphi\urcorner)$$
$$PA_d + \forall x(Thm_{PA}(x) \rightarrow Vr(x)) \vdash Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$
$$\text{D'où } PA_d + \forall x(Thm_{PA}(x) \rightarrow Vr(x)) \vdash Thm_{PA}(\ulcorner\varphi\urcorner) \rightarrow \varphi$$

Et donc $PA_d + Ref(PA)$ prouve (toutes les instances de) Rfn(PA). Notez qu'il n'est pas besoin d'employer l'induction sur des formules contenant le prédicat de vérité.

Pour obtenir le principe de réflexion uniforme à partir de $Ref(PA)$, une extension aléthique un peu plus forte que la simple décitation PA_d est nécessaire. Il faut en effet pouvoir faire appel à la version uniforme des **T**-équivalences , *i.e.* se placer (au moins) dans l'extension aléthique décitationnelle uniforme $PA_{d.u.}$

$$PA_{d.u.} + Ref(PA) \vdash \text{RFN(PA)} \text{ [183]}$$

181. En s'appuyant sur le fait que PA est une théorie Σ^0_1-correcte, ce résultat peut s'obtenir comme un corollaire relativement direct d'un autre théorème important dû à Kreisel et Lévy (1968), :

Théorème. Théorème de non-bornage (Kreisel et Lévy, 1968)
Soit T une théorie récursivement axiomatisée. Alors

1. Rfn(T) n'est contenu dans aucune extension *finie* consistante de T.
2. RFN(T) n'est contenu dans aucune extension consistante de T de *complexité arithmétique bornée.*

182. Néanmoins, on peut aussi voir ce résultat comme une conséquence d'un résultat plus fort dû à Halbach (1999b, proposition 2 p . 13). Voyez ci-dessous.

Voilà donc pour des premiers liens entre l'axiome unique du principe global de réflexion et les schémas d'axiomes formant la réflexion locale ou uniforme : les seconds peuvent être (très) facilement dérivés du premier. Mais on constate toutefois que ces dérivations dépendent crucialement de la présence d'axiomes aléthiques auxiliaires (sous la forme **T**-équivalences décitationnelles plus ou moins fortes). Y a-t-il un rapport d'implication inverse entre réflexion locale ou uniforme et le principe global de réflexion ? Dans la mesure où $Ref(PA)$ n'est même pas exprimable dans le langage $\mathcal{L}_{PA}$, il n'y a guère de sens à se demander de manière directe si une forme de réciproque des résultats ci-dessus vaut, au sens où $PA + \text{Rfn(PA)}$ (resp. $PA + \text{RFN(PA)}$) prouverait $Ref(PA)$. Toutefois, si on se limite aux énoncés de $\mathcal{L}_{PA}$ on a un résultat intéressant dû une fois encore à Halbach. Halbach (1999b) établit que $PA + \text{Rfn(PA)}$ a exactement les mêmes conséquences *arithmétiques* que $PA_d + Ref(PA)$ —et de même pour $PA + \text{RFN(PA)}$ et $PA_{d.u.} + Ref(PA)$. En effet, plus généralement,

Proposition. *(*Halbach*, 1999b, proposition 2 p. 13)*

Soit T une théorie exprimée dans $\mathcal{L}_{PA}$ et capable de représenter sa propre syntaxe. Soit $\Phi(x)$ une formule de $\mathcal{L}_{PA}$ contenant x pour seule variable libre. Alors,

$$\Big(T_d^{+Ind} \cup \forall x(\Phi(x) \to Vr(x))\Big)^{\vdash} \cap \mathcal{L}_{PA} = \Big(T \cup \{\Phi(\ulcorner\varphi\urcorner) \to \varphi \mid \varphi \in \mathcal{L}_{PA}\}\Big)^{\vdash}$$ [184]

En particulier, on a donc

Corollaire. *Pour toute énoncé $\varphi \in \mathcal{L}_{PA}$,*

$$PA_d^{+Ind} + Ref(PA) \vdash \varphi \textit{ ssi } PA + \text{Rfn(PA)} \vdash \varphi$$
$$PA_{d.u.}^{+Ind} + Ref(PA) \vdash \varphi \textit{ ssi } PA + \text{RFN(PA)} \vdash \varphi$$ [185]

En d'autres termes, $PA_d^{+Ind} + Ref(PA)$ (resp. $PA_{d.u.}^{+Ind} + Ref(PA)$) est une extension

183. Pour mémoire,

$$PA_{d.u.} = PA \cup \{\forall x_1, \ldots, \forall x_n \big(Vr(\ulcorner\varphi(\mathring{x_1}, \ldots, \mathring{x_n})\urcorner) \leftrightarrow \varphi(x_1, \ldots, x_n)\big) \mid \varphi(\vec{x_i}) \in Form(\mathcal{L}_{PA})\}$$

Comme dans le cas de la réflexion locale, il existe une démonstration directe et quasi triviale de ce résultat, analogue à la précédente. De même, on peut aussi voir ce résultat concernant la réflexion uniforme comme une conséquence d'un résultat plus fort dû à Halbach (1999b, proposition 2 p . 13). Pour plus de détails voyez Halbach (2014).

184. où $X^{\vdash}$ désigne la clôture déductive de X.

185. Le cas de la décitation uniforme ne découle pas directement de la proposition ci-dessus. Néanmoins, selon Halbach (2014, chapitre 22, p. 324, note 2) la démonstration peut aisément être adaptée pour obtenir le cas reliant la décitation uniforme et le principe de réflexion uniforme.

conservative de $PA + \text{Rfn(PA)}$ (resp. $PA + \text{RFN(PA)}$). [186]

Par ailleurs, nous avons déjà vu en détails que l'extension aléthique tarskienne munie de l'induction élargie PA^{+ind}_{Tar} prouve $Ref(PA)$, le principe global de réflexion. Dans cette extension aléthique, il n'est donc pas nécessaire d'ajouter $Ref(PA)$ comme un énoncé indépendant. Étant donné qu'en outre PA^{+ind}_{Tar} permet de dériver toutes les instances décitationnelles uniformes (*i.e.* $PA_d \subset PA_{d.u.} \subset PA^{+ind}_{Tar}$), il suit immédiatement que

$$\begin{aligned} \left(PA^{+ind}_{Tar} \vdash Ref(PA)\right) &\text{ (déjà vu)} \\ \text{D'où, } \; PA^{+ind}_{Tar} &\vdash \text{RFN(PA)} \\ \text{Et } \textit{a fortiori}, \; PA^{+ind}_{Tar} &\vdash \text{Rfn(PA)} \end{aligned}$$

Pour en finir avec ces rappels techniques, signalons enfin que PA^{+ind}_{Tar} n'est pas conservative sur $PA + \text{RFN(PA)}$. Comme le note Halbach (2014, p. 325), une analyse preuve-théorique précise de PA^{+ind}_{Tar} montre que c'est une extension beaucoup plus forte que celle obtenue en ajoutant simplement à PA le schéma de réflexion uniforme. [187] De là, il suit

186. Notez au passage que

1. premièrement, couplé avec les resultats précédents ceci a pour conséquence que
   (a) $PA^{+Ind}_{d} + Ref(PA) \nvdash PA + \text{RFN(PA)}$
   (b) et donc que $\left(PA^{+Ind}_{d.u.} + Ref(PA)\right) \restriction_{\mathcal{L}_{PA}}$ est une extension *stricte* de $\left(PA^{+Ind}_{d} + Ref(PA)\right) \restriction_{\mathcal{L}_{PA}}$

   Ainsi, en présence de $Ref(PA)$ (et de l'induction étendue), le passage de la décitation stricte à la décitation uniforme *a un impact* sur la force *arithmétique* de l'extension aléthique considérée.
2. deuxièmement, puisqu'on a déjà établi

   $$PA_d + Ref(PA) \vdash \text{Rfn(PA)} \text{ et } PA_{d.u.} + Ref(PA) \vdash \text{RFN(PA)}$$

   le corollaire a pour conséquence immédiate que Pour toute énoncé $\varphi \in \mathcal{L}_{PA}$,

   $$\begin{aligned} PA^{+Ind}_{d} + Ref(PA) \vdash \varphi \text{ ssi } \; PA + \text{Rfn(PA)} \vdash \varphi \text{ ssi } \; PA_d + Ref(PA) \vdash \varphi \\ PA^{+Ind}_{d.u.} + Ref(PA) \vdash \varphi \text{ ssi } PA + \text{RFN(PA)} \vdash \varphi \text{ ssi } PA_{d.u.} + Ref(PA) \vdash \varphi \end{aligned}$$

   Ce qui implique aussitôt que $PA^{+Ind}_{d} + Ref(PA)$ et $PA_d + Ref(PA)$ prouvent exactement les mêmes énoncés de $\mathcal{L}_{PA}$ tout comme $PA^{+Ind}_{d.u.} + Ref(PA)$ et $PA_{d.u.} + Ref(PA)$. Autrement dit encore, en présence du schéma de réflexion global $Ref(PA)$ et d'un prédicat de vérité axiomatisé par une théorie décitationnelle (uniforme ou simple), l'extension de l'induction au langage $\mathcal{L}_{Vr}$ n'a *strictement aucun* impact sur la capacité de la théorie étendue à prouver des énoncés de $\mathcal{L}_{PA}$, c'est-à-dire des énoncés purement arithmétiques.

Ces résultats semblent confirmer une fois encore que les axiomes pour la vérité tiennent bien un rôle crucial dans l'augmentation de la force *arithmétique* des extension aléthiques considérées. Une pierre de plus dans le jardin de Field (1999).

187. Voir Halbach (2014, p. 325-326). Pour prouver que PA^{+ind}_{Tar} est une extension stricte de $PA +$

également que PA^{+ind}_{Tar} est une extension stricte de $PA_{d.u.} + Ref(PA)$. Ainsi, accepter les clauses tarskiennes pour PA (et l'induction sur les énoncés contenant du vocabulaire sémantique) c'est s'engager à beaucoup plus qu'à simplement accepter l'énoncé de la correction de PA sous la forme d'un principe de réflexion global pour PA (ou uniforme, ou local d'ailleurs).

Au total, si on note $X \subsetneq Y$ le fait qu' Y est une extension stricte (*i.e.* non-conservative) de X et $X =_{\mathcal{L}_{PA}} Y$ le fait que X et Y prouvent *exactement* les mêmes théorèmes du langage $\mathcal{L}_{PA}$, nous avons donc le diagramme suivant qui résume les résultats ci-dessus :

$$\begin{array}{ccccc}
PA + Con(PA) & \subsetneq & PA + \mathrm{Rfn(PA)} & \subsetneq & PA + \mathrm{RFN(PA)} \\
 & & \|_{\mathcal{L}_{PA}} & & \|_{\mathcal{L}_{PA}} \\
 & & PA^{+ind}_{d.} + Ref(PA) & \subsetneq & PA^{+Ind}_{d.u.} + Ref(PA) \subsetneq PA^{+Ind}_{Tar} \\
 & & \|_{\mathcal{L}_{PA}} & & \|_{\mathcal{L}_{PA}} \\
 & & PA_{d.} + Ref(PA) & \subsetneq & PA_{d.u.} + Ref(PA)
\end{array}$$

Ces rappels techniques étant effectués, revenons en à présent à la discussion philosophique du déflationnisme.

4.3.2 La querelle opposant Ketland et Tennant

4.3.2.1 Nul besoin de Vérité : Tennant (2002) et les schémas de réflexion

La position développée par Tennant (2002) comporte principalement deux aspects, complémentaires l'un de l'autre. Le premier, que nous avons déjà évoqué à plusieurs reprises, consiste à proposer au nom du déflationnisme de renoncer à la contrainte de réflexivité. Voici comment Tennant (2002) lui-même présente les choses. Il résume d'abord la situation logique comme ceci [188] : pour toute théorie de base du premier ordre T qui contient « assez » d'arithmétique, il n'est pas possible de construire $V(T)$, une extension consistante de cette théorie T, vérifiant les trois propriétés suivantes :

RFN(PA), Halbach montre qu'on peut dériver dans PA^{+ind}_{Tar} un principe de réflexion uniforme pour $PA + \mathrm{RFN(PA)}$ (et donc aussi un énoncé du type $Con(PA + \mathrm{RFN(PA)})$ qui bien évidemment n'est pas dérivable dans $PA + \mathrm{RFN(PA)}$ alors même que c'est un énoncé de $\mathcal{L}_{PA}$).

188. *cf.* Tennant (2002, p. 566).

(a) $V(T)$ est une extension conservative de T.

(b) $V(T)$ prouve « tous les théorèmes de T sont vrais ».

(c) $V(T)$ satisfait le critère de la CONVENTION **T**, *i.e.* elle prouve chaque **T**-équivalence $Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$, pour chaque énoncé φ du langage de T.

Selon Tennant, (a) et (c) ci-dessus ne sont guère contestables et s'imposent pour ainsi dire au déflationniste. Pour sortir du piège de Shapiro et Ketland, il faut donc renoncer à (b). Il écrit ainsi :

> Pour quelqu'un qui considère effectivement l'hypothèse de conservativité comme essentielle au déflationnisme, il ne reste qu'une seule porte de sortie : abandonner l'affirmation dans [l'extension aléthique pour T] [189] que tous les théorèmes de T sont vrais. Shapiro semble penser que cela serait une terrible conséquence pour le déflationnisme, révélant son inadéquation en tant que théorie de la vérité. [...] Mais l'abandon de l'affirmation, dans [l'extension aléthique], que tous les théorèmes de T sont vrais serait-elle une si terrible conséquence pour le déflationnisme ? [...]
>
> La question qui émerge est la suivante : *devrions-nous nous préoccuper d'énoncer la correction sous une forme qui se réfère explicitement à la préservation de la vérité* ? Peut-être que tout ce à quoi parvient l'aporie de Shapiro, c'est à nous faire apercevoir que la réponse à cette dernière question devrait être négative. [...] On nous demande de croire que l'affirmation à venir dans $[V(T)]$
>
> Tous les théorèmes de $[T]$ sont vrais
>
> est la seule [...] manière d'exprimer notre conviction reflexive concernant la « correction de T ». *Mais est-ce là l'unique manière d'exprimer cette conviction ? Au nom du déflationnisme nous nous risquons à suggérer que non.* TENNANT (2002, p. 568-569, nous soulignons)

Mais bien entendu, il ne suffit pas de stipuler qu'on renonce à construire une axiomatisation de la vérité satisfaisant la contrainte de réflexivité. Encore faut-il montrer qu'une telle position est justifiable.

189. Nous modifions ici la notation de TENNANT (2002), qui note S la théorie arithmétique de base et S^* ses extensions.

C'est justement le second aspect du travail de Tennant (2002). Pour légitimer l'abandon de la réflexivité aléthique,[190] Tennant se propose de montrer qu'il existe d'autres moyens, non sémantiques, d'articuler notre « conviction réflexive » concernant la correction de T. Plus précisément, il s'agit de montrer que les arguments sémantiques « standard » avancés par Shapiro et Ketland peuvent être reconstruits, ou peut-être remplacés, par des arguments ne s'appuyant pas sur la notion de vérité. Ainsi Tennant déclare :

> Ma réplique à la fois à Shapiro et à Ketland reviendra à ceci : le déflationniste possèdes des moyens « philosophiquement modestes » de mener à bien le soi-disant argument « sémantique » établissant la « vérité » de l'énoncé de Gödel. En fait, cet argument peut être directement et fidèlement enrégimenté dans des termes dont on peut immédiatement voir qu'ils sont déflationnistiquement licites. Tennant (2002, p. 553)

Pour être absolument certain que cette reformulation ou enrégimentation de l'argument sémantique obéira bien aux canons déflationnistes et qu'on n'y fera pas subrepticement usage d'une notion substantielle ou explicative de vérité, Tennant impose qu'aucun emploi ou mention du prédicat de vérité n'apparaisse dans la dérivation de G qu'il va proposer.[191] En outre, Tennant ajoute que la « structure profonde », ou la « ligne de pensée » de l'argument sémantique classique évoqué par Ketland et Shapiro sera préservées par sa reformulation répondant aux canons déflationnistes, au point qu'elle

> [sera], et ce de manière manifeste, un homologue formel de l'argument sémantique une fois celui-ci convenablement compris. (Tennant, 2002, p. 557)

Les intentions de Tennant sont donc claires. Malgré cela, un phénomène assez étrange se produit : lorsqu'il expose l'argument sémantique qu'il va s'agir de « déflater », Tennant ne reprend pas l'argument donné par Ketland (1999) et Shapiro (1998b) mais reproduit un argument un peu différent, quoiqu'il soit lui aussi très répandu dans la littérature logique. Voici donc la version *verbatim* donnée par Tennant lui-même de l'argument sémantique que Tennant entend recomposer à partir de ressources admissibles d'un point de vue strictement déflationniste :

190. *i.e.* la capacité à prouver dans $V(T)$ que tous les théorèmes de T sont vrais.
191. Tennant s'en explique de la manière suivante (Tennant, 2002, p. 562) :

> La raison pour laquelle nous entreprenons de nous interdire le prédicat de vérité est pour rendre évident que justice déflationniste a été rendue vis-à-vis de l'argument sémantique.

Argument sémantique$^{(*)}$ [selon Tennant (2002)] pour la vérité de l'énoncé de Gödel :

> G est un énoncé universellement quantifié (en l'occurrence, d'un type à la Goldbach, c'est-à-dire une quantification universelle d'un prédicat primitif récursif). Chaque instance numérique de ce prédicat est prouvable dans T. [...] Une preuve dans T garantit la *vérité*. D'où, chaque instance numérique de G est *vraie*. Donc, puisque G est simplement la quantification universelle de ces instances numériques, elle aussi doit être *vraie*. Tennant (2002, p. 556)

On pourra constater — et ce point est en fait d'une importance cruciale— que cet argument [192] est quelque peu différent de celui évoqué par Ketland et Shapiro et que nous avons largement exposé en détails dans ce qui précède. Ici, point de raisonnement inductif sur la longueur des preuves dans T permettant d'établir un principe de réflexion (global) pour T ; mais plutôt le passage de la vérité de toutes les instances numériques d'une formule primitive-récursive à la vérité de la quantification universelle de cette formule, la vérité de chaque instance numérique étant assurée par le fait qu'elles sont chacune prouvables dans T et que *les preuves dans T garantissent la vérité*. Pour distinguer cet argument de celui de Ketland et Shapiro, nous le noterons : « argument sémantique* ».

Pour formaliser *cet* argument sémantique* en des termes déflationnistiquement autorisés, Tennant propose de recourir aux principes de réflexion, et plus particulièrement aux versions schématiques de ces principes, exprimées dans le langage de la théorie de base, *i.e.* sans employer de prédicat de vérité. Tennant examine plusieurs versions de schémas de réflexion et fait porter son choix sur une version restreinte du principe de réflexion uniforme [193] :

$\mathrm{RFN}_{\mathrm{P.R.}}(\mathrm{T})$ $\quad \forall x Thm_T(\ulcorner \varphi(\dot{x}) \urcorner) \rightarrow \forall x \varphi(x)$,

où $\varphi(x)$ est une formule de $\mathcal{L}_T$ (représentant une relation) primitive récursive (d'où le suffixe P.R.).

Dans $T + \mathrm{RFN}_{\mathrm{P.R.}}(\mathrm{T})$, *i.e.* la théorie de base étendue au moyen de ce schéma de réflexion uniforme pour les seules formules primitives récursives, l'énoncé de Gödel G_T pour T, indécidable dans T elle-même, devient dérivable. Tennant (2002, section 9, p. 577) en donne une démonstration qui revient à peu près à ceci :

192. Cet argument est en réalité directement repris de Dummett (1963), comme l'indique Tennant lui-même.
193. *cf.* Tennant (2002, p. 572-573).

Rappelons que G_T est obtenue au moyen d'une construction de point fixe pour la formule $\neg\exists x Dem_T(x, y)$ et vérifie :

$$(Diag) : T \vdash G_T \leftrightarrow \neg\exists x Dem_T(x, \ulcorner G_T \urcorner)$$

où $Dem_T(x, y)$ est une formule représentant (dans T) la relation *primitive récursive* x est (le code d') une preuve de la formule (codée par) y. D'après le premier théorème d'incomplétude, sous l'hypothèse que T est cohérente, $T \nvdash G_T$. Par conséquent, par représentabilité de la relation de « x est une preuve de y dans T », pour tout entier n, $T \vdash \neg Dem_T(\overline{n}, \ulcorner G_T \urcorner)$. Tennant raisonne alors comme suit :

1. $T \vdash \neg Dem_T(\overline{n}, \ulcorner G \urcorner)$	pour tout entier n
2. $T \vdash Thm_T(\ulcorner \neg Dem_T(\overline{n}, \ulcorner G \urcorner) \urcorner)$	pour tout entier n
3. $T \vdash \forall x Thm_T(\ulcorner \neg Dem_T(\dot{x}, \ulcorner G \urcorner) \urcorner)$	
4. $T + \mathrm{RFN}_{\mathrm{P.R.}}(\mathrm{T}) \vdash \forall x Thm_T(\ulcorner \neg Dem_T(\dot{x}, \ulcorner G \urcorner) \urcorner) \rightarrow \forall x \neg Dem_T(x, \ulcorner G \urcorner)$	par réflexion uniforme [194]
5. $T + \mathrm{RFN}_{\mathrm{P.R.}}(\mathrm{T}) \vdash \forall x \neg Dem_T(x, \ulcorner G \urcorner)$	par 3., 4., modus ponens
6. $T + \mathrm{RFN}_{\mathrm{P.R.}}(\mathrm{T}) \vdash G_T$	par $(Diag)$ [195]

Selon Tennant, cette démonstration constitue un « analogue formel » de l'argument sémantique* que nous avons retranscrit à la page précédente (voir p. 349). La forme logique de l'argument est en effet en partie préservée :

> G_T est (équivalente à) un énoncé universellement quantifié (à savoir, $\forall x \neg Dem_T(x, \ulcorner G_T \urcorner)$). Chaque instance numérique de cet énoncé universel, *i.e.* chaque instance $\neg Dem_T(\overline{n}, \ulcorner G_T \urcorner)$, est démontrable dans T, ce dernier fait étant lui-même démontrable dans T, au sens où $T \vdash Thm_T(\ulcorner \neg Dem_T(\overline{n}, \ulcorner G \urcorner) \urcorner)$. Le principe de réflexion uniforme permet alors de passer de la prouvabilité de chaque instance numérique de G_T (ou plus précisément de $\neg Dem_T(\overline{n}, \ulcorner G_T \urcorner)$) à une preuve de la quantification universelle elle-même (*i.e.* $\forall x \neg Dem_T(x, \ulcorner G_T \urcorner)$) et donc de G_T.

Dans la mesure où $\neg Dem_T(x, \ulcorner G_T \urcorner)$ est une formule (représentant une relation) primitive récursive, on peut limiter l'extension de T à un principe de réflexion uniforme

194. Notez que la réflexion uniforme limitée aux formules P.R. suffit puisque $\neg Dem_T(x, \ulcorner G \urcorner)$ est primitive récursive.

195. Et l'équivalence triviale $T \vdash \neg\exists x Dem_T(x, \ulcorner G_T \urcorner) \leftrightarrow \forall x \neg Dem_T(x, \ulcorner G_T \urcorner)$.

restreint à la classe des ces seules formules de $\mathcal{L}_T$. Tennant mentionne d'ailleurs que ce principe de réflexion est exactement de la bonne force logique au sens où il est en fait équivalent à G_T (et à $Con(T)$) modulo T : il est donc minimalement suffisant pour dériver G_T (ou $Con(T)$). D'après Tennant, ce principe de réflexion est donc à la fois de la bonne *force* logique (puisqu'il permet de prouver exactement ce qu'il faut sans nous obliger à nous engager envers une théorie inutilement forte) et de la bonne *forme* logique puisqu'il permet de construire un homologue respectant la « structure profonde » ou « la ligne de pensée » de l'argument sémantique*.

Pour Tennant, la démonstration ci-dessus illustre donc le fait que le déflationniste trouve à sa disposition des moyens « philosophiquement modestes » et parfaitement acceptables selon ses critères, de mener à bien un « analogue formel » de l'argument sémantique*, lequel argument une fois ainsi enrégimenté n'a alors plus rien de sémantique puisqu'aucune notion de cet ordre n'y ait ne serait-ce que mentionnée. Et cette reconstruction montre par conséquent qu'on peut fournir au déflationniste les moyens d'expliquer que G_T ou $Con(T)$ sont légitimement assertables (Tennant évite soigneusement ici d'employer le mot « vrai ») sans qu'il soit besoin de recourir à une théorie non-conservative de la vérité.

Il est alors loisible au déflationniste d'abandonner la contrainte de réflexivité.

4.3.2.2 La réaction de Ketland (2005) et le problème de la justification des schémas de réflexion

La plaidoirie déployée par Tennant en faveur du déflationnisme a rapidement suscité une réplique de la part de Ketland (2005), laquelle a à son tour elle-même donné lieu à une réponse de Tennant (2005). Nous avons souligné que l' « argument sémantique* » mis en avant par Tennant (2002, p. 556), et dont ce dernier entendait proposer une « reconstruction déflatée », était différent dans sa structure de celui évoqué par Ketland (1999) ou Shapiro (1998b). Ce point qui pourrait paraître anodin, simple variation sur le même thème, recouvre en fait la pierre d'achoppement fondamentale dans la querelle qui oppose Ketland (et Shapiro) à Tennant.

Selon Ketland (2005) en effet, le raisonnement développé par Tennant repose sur une mauvaise compréhension de ce en quoi consiste la contrainte de réflexivité. [196]

196. *Cf.* (Ketland, 2005, p. 76) :

> « Tennant mésinterprète les points centraux de l'argument que Shapiro et moi-même avons donné. »

D'après lui, la contrainte de réflexivité concernait

> « la possibilité de fournir une *justification* [sous la forme d'une preuve] du principe de réflexion : "tous les théorèmes de [T] sont vrais" » (Ketland, 2005, p. 76)

une fois qu'on a déjà accepté une théorie mathématique de base [T]. Bien évidemment, une telle justification fournit aussi correlativement une justification de principes de réflexion plus faibles tels que la réflexion locale ou uniforme (Rfn ou RFN) puisque ces derniers peuvent être dérivés du principe de réflexion global. [197] Mais le point crucial est que dans l'argument sémantique évoqué par Ketland et Shapiro dans leurs articles respectifs, le principe de réflexion se trouve à la conclusion [198] d'une dérivation, laquelle peut être formalisée dans une extension aléthique tarskienne. [199] Cette capacité à prouver le principe (global) de réflexion à partir d'axiomes aléthiques fait qu'une théorie réflexive de la vérité est censée fournir une *justification* et une *explication* des raisons pour lesquelles quelqu'un qui accepte une théorie T devrait également accepter les instances de divers schémas du type $Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$. En ce sens, pour Ketland la justification des principes de réflexion mobilise la notion de vérité. À l'inverse, dans l'argument à la Dummett retracé par Tennant, tout comme dans la reconstruction qu'il en donne, le principe de réflexion (en l'occurrence le principe restreint $\mathrm{RFN}_{\mathrm{P.R.}}(\mathrm{T})$) est employé comme un postulat ou un axiome. Par conséquent, cet argument s'*appuie* sur un principe de réflexion et ne peut donc pas à son tour être considéré comme fournissant une *justification* du principe de réflexion lui-même. Aux yeux de Ketland (2005), la proposition de Tennant (2002) revient donc à postuler sans autre forme de procès ce qu'il s'agit de démontrer, ou du moins de justifier.

De façon plus détaillée, au cours de sa réaction au plaidoyer de Tennant, Ketland approfondit et précise sa propre position pour en accentuer le contraste avec celle défendue par Tennant. Voici ce qu'il développe : Ketland rappelle qu'il existe une différence logique manifeste entre accepter une théorie T et tous ses théorèmes, et accepter un principe de réflexion sur T (quelle que soit sa formulation). C'est en somme la leçon de la non-conservativité de RFN(T)/Rfn(T)/$Ref(T)$ sur T. Pour autant, quiconque accepte

.

197. Modulo certains principes aléthiques tels que la décitation ou la décitation uniforme (voyez sur ce point nos rappels techniques de la section 4.3.1).

198. conclusion qui au demeurant peut n'être qu'une étape intermédiaire si on poursuit le raisonnement pour obtenir une preuve de $Con(T)$ ou de G_T.

199. au prix bien entendu de la non-conservativité.

une théorie T devrait aussi accepter certaines propositions qui ne sont pas directement conséquences logiques de T. S'appuyant sur les travaux de Feferman, Ketland introduit ainsi une notion d'

> **Obligation épistémique conditionnelle**
>
> *Si* quelqu'un accepte une théorie mathématique de base S, alors il est tenu d'accepter un certain nombre d'*affirmations supplémentaires* dans le langage la théorie de base (Ketland, 2005, p. 79, italiques de l'auteur)

Parmi ces affirmations qui semblent s'imposer à quiconque a accepté T, on retrouve certains principes de réflexion et certains énoncés indécidables dans T tels que G_T ou $Con(T)$. Pour délimiter plus précisément ce que sont ces affirmations, Ketland renvoie aux travaux de Feferman :

> [...] quels énoncés du langage de base $[\mathcal{L}_T]$ de $[T]$ [200] [...] devraient être acceptés si on a accepté les axiomes de base et les règles de $[T]$? La réponse est donnée sous la forme d'une théorie ordinaire $[ExtAl(T)]$ [201] formulée dans un langage $[\mathcal{L}_T(Vr, F)]$. [...] Ainsi, par exemple, on peut raisonner dans $[ExtAl(PA)]$ par induction sur la vérité d'énoncés qui contiennent la notion de vérité, et parvenir de cette manière à des énoncés de la forme : $\forall x(Thm_{PA}(x) \rightarrow Vr(x))$, et en itérant ce genre d'arguments dériver des principes de réflexion itérés pour l'arithmétique. (Feferman, 1991, p. 2) (cité dans Ketland (2005, p. 80))

Pour Ketland, le raisonnement sémantique formalisé dans une extension aléthique réflexive est donc une manière d'expliquer ou de mettre au jour le raisonnement réflexif qui est à l'œuvre derrière l'obligation épistémique conditionnelle :

> Les énoncés envers lesquels nous sommes engagés en acceptant une théorie

200. Notation modifiée.

201. Ici Feferman introduit une extension aléthique (notée $Ref(T)$ dans le texte original, notation que nous ne reprenons pas pour éviter toute confusion avec la principe global de réflexion) exprimée dans un langage étendu au moyen d'un prédicat de vérité et d'un prédicat de fausseté, et qui axiomatise une notion non-typée de vérité (*i.e.* pouvant s'appliquer à des énoncés contenant déjà eux-mêmes des notions sémantiques). Cette axiomatisation se situe dans lignée de celle proposée par Kripke (1975) et est usuellement connue sous le nom d'axiomatisation KF ou Kripke-Feferman. Comme nous nous limitons ici au cas d'un prédicat de vérité typé, une analyse de ce système aléthique déborde le cadre de ce travail (pour plus d'informations techniques sur ce sujet voyez par exemple Halbach (2014, chapitre 15-16)). Notez cependant que l'extension aléthique ($KF(PA)+$ induction étendue au vocabulaire sémantique, *i.e.* PA^{+Ind}_{KF}) est plus forte que l'extension tarskienne PA^{+ind}_{Tar} et que le raisonnement inductif sur la longueur des preuves permettant d'établir le principe global de réflexion $Ref(PA)$ y est largement fomalisable.

> de base $[T]$ pourrait être appelés les *conséquences réflexives*[202] de $[T]$. Ces conséquences réflexives incluent l'énoncé de Gödel $[G_T]$ de la théorie, l'énoncé de la consistance $[Con(T)]$, plus les schémas de réflexion local et uniforme. Étendre une théorie de base $[T]$ avec des axiomes pour la vérité est ainsi une manière d'*extraire déductivement*[203] ces conséquences réflexives. (Ketland, 2005, p. 80, note 8)

Par « extraire déductivement », il faut visiblement comprendre ici « donner une preuve en bonne et due forme et, ce faisant, fournir une justification ».

Parallèlement, Ketland insiste sur le fait que dans l'argument évoqué par Tennant (2002), l'une des hypothèses cruciales est que *les preuves dans T garantissent la vérité.*[204] Cette hypothèse de la *correction* de T exprimée ici en langage naturel, est typiquement formalisée par un principe de réflexion : soit par le principe global si on veut rester le plus proche possible de la formulation informelle employant le prédicat de vérité, soit, comme c'est le cas dans la reconstruction déflatée proposée par Tennant, au moyen d'un schéma d'axiomes de réflexion locale ou uniforme. Selon Ketland (2005, p. 81), cette hypothèse est simplement postulée sans argument par Tennant, tandis que Shapiro, Feferman et lui-même ne se contentent pas de la supposer, mais la démontrent :

> nous fournissons un argument séparé crucial pour la *correction* de $[T]$.[205] À l'inverse, Tennant ne donne aucune justification pour la prémisse cruciale, selon laquelle la théorie $[T]$ est *correcte*. (Ketland, 2005, p. 82)

Au yeux de Ketland, Tennant est donc loin d'avoir montré que les déflationnistes possèdent les moyens de contourner ou dépasser de manière satisfaisante la contrainte de réflexivité :

> [...] Tennant n'est pas parvenu à montrer que de telles stratégies sont « disponibles », sauf si *supposer quelque chose sans argument* le rend « dis-

202. italiques de l'auteur.
203. nous soulignons.
204. Voyez l'extrait cité page 349 :

> **Argument sémantique* [selon Tennant (2002)] pour la vérité de l'énoncé de Gödel :**
> Chaque instance numérique de ce prédicat est prouvable dans T. [...] **Une preuve dans T garantit la *vérité***. D'où, chaque instance numérique de G est *vraie*. [...] Tennant (2002, p. 556)

205. L'argument auquel il est fait référence ici est évidemment la preuve par induction sur la longueur des preuves du schéma de réflexion (global) donnée dans une extension aléthique réflexive.

ponible ». (Ketland, 2005, p. 85)

Ketland ne veut pas exclure *a priori* qu'il soit peut-être possible pour le déflationniste de proposer une autre justification des principes de réflexion, qui ne soit pas basée sur une théorie de la vérité. [206] Mais à ses yeux, Tennant est pour le moment bien loin d'avoir fourni une telle explication de remplacement. Et Ketland conclut donc :

> [t]out en rejetant l'explication des principes de réflexion appuyée sur une théorie de la vérité donnée par Feferman, Shapiro et moi-même, Tennant se contente de *postuler* un principe de réflexion, sans fournir de justification à son appui, illustrant par là ce que Bertrand Russell a vait jadis appelé « les avantages du vol sur le labeur honnête ». (Ketland, 2005, p. 87)

Il faut bien reconnaitre que, principalement occupé à donner une reformulation de l'argument sémantique* à partir d'un schéma de réflexion dans lequel le prédicat de vérité n'apparaît pas, Tennant ne s'est guère attardé dans son article de 2002 sur la question de savoir comment les principes de réflexion eux-mêmes peuvent être justifiés. Le seul passage de Tennant (2002) dans lequel la critique de Ketland (2005) est en quelque sorte anticipée se trouve vers la fin de l'article :

> Un anti-déflationniste pourrait objecter ici que, dans la version du processus de réflexion ou de justification donnée par Shapiro, la conclusion de la correction de $[T]$ était obtenue *seulement après un argument* dont les étapes détaillées n'ont pas été préservées par notre déflationniste qui adopte simplement en remplacement le principe de correction ci-dessus. [207] Cet argument se référait à la vérité des axiomes de $[T]$ et au caractère de préservation de la vérité des règles de $[T]$. Le déflationniste possède à portée de main une réponse à cette objection. Prenons simplement des règles de remplacement (qui mènent de prémisses concernant la prouvabilité dans $[T]$ à des conclusions formulées comme de simples assertions et non pas comme des prédications impliquant la vérité) [208] correspondant à chacun des axiomes et des règles

206. *Cf.* Ketland (2005, p. 80 et 87). On peut noter ici une inflexion de la position de Ketland par rapport à son article de 1999. Dans Ketland (1999), il semble véritablement défendre une forme d'inspensabilité de la justification sémantique, au sens où la preuve par la vérité des schémas de réflexion serait *la* justification incontournable ou la seule possible. Dans Ketland (2005) à l'inverse, il laisse ouverte la possibilité qu'une autre explication (non sémantique) soit possible, mais considère qu'elle reste à découvrir et que ni Tennant ni les déflationnistes ne l'ont pour l'instant produite.

207. Tennant fait ici référence à un schéma de réflexion uniforme formulé sans prédicat de vérité.

208. Tennant passe très rapidement sur ce qu'il veut dire par là. Sans doute la question de la reformu-

> d'inférence à partir desquels les preuves dans $[T]$ sont construites. Dès lors, le principe de correction devrait devenir dérivable dans le système [étendu T^*], et le déflationniste aura imité la structure du processus de justification de Shapiro. (Tennant, 2002, p. 575)

Face aux critiques pressantes de Ketland (2005), Tennant (2005) donne plus de détails sur la façon dont, selon lui, on pourrait justifier les principes de réflexion sans s'appuyer sur la notion de vérité. Curieusement, il n'est plus question ici de reproduire les étapes de la preuve sémantique de (Ref) dans PA^{+ind}_{Tar}. Tennant semble plutôt vouloir

---

lation précise de l'argument inductif de Shapiro (et Ketland) sur la longueur des preuves mériterait plus qu'une simple remarque entre parenthèses. Tel que nous comprenons ce passage de Tennant (2002), nous serions tentés de dire que, de même que le principe de réflexion global formulé avec un prédicat de vérité :

$$\text{Ref(T)} : \forall x(Thm_T(x) \rightarrow Vr(x))$$

peut selon Tennant être remplacé par le schéma

$$\text{Rfn(T)} : Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi, \varphi \in \mathcal{L}_T$$

de même, l'énoncé « tous les axiomes de T sont vrais » :

$$AxVr(T) : \forall x(Ax_T(x) \rightarrow Vr(x))$$

peut sans doute être remplacé par le schéma

$$Ax_T(\ulcorner\varphi\urcorner) \rightarrow \varphi, \varphi \in \mathcal{L}_T$$

Le cas de l'énoncé « les règles d'inférence de T préservent la vérité » est un peu moins évident. Néanmoins,

$$InfVr(T) : \forall x \forall y \forall x((Inf_T(x, y, z) \wedge Vr(y) \wedge Vr(z)) \rightarrow Vr(x))$$

où $(Inf_T(x, y, z)$ signifie que la formule codée par x est obtenue au moyen d'une règle d'inférence de T à partir des formules codées par y et z (typiquement par application d'un modus ponens), peut, peut-être, être remplacé par le schéma

$$(Inf_T(\ulcorner\varphi\urcorner, \ulcorner\psi_1\urcorner, \ulcorner\psi_2\urcorner) \wedge \psi_1 \wedge \psi_2) \rightarrow \varphi, \text{où } \psi_1, \psi_2, \varphi \in \mathcal{L}_T$$

Ces versions schématiques, expurgées de notions sémantiques, de « tous les axiomes sont vrais », et de « les règles d'inférence préservent la vérité » suffisent-elles pour « imiter » ou « reformuler » un argument inductif sur la longueur des preuves permettant d'établir un principe de réflexion uniforme ou local (*i.e.* d'obtenir en conclusion une version non aléthique de « tous les théorèmes de T sont vrai ») ? À vrai dire, cela nous semble loin d'être aussi évident que Tennant le laisse entendre. En particulier, la preuve semantique menée dans PA^{+ind}_{Tar} du principe de réflexion global $Ref(T)$ comporte de nombreuses quantifications sur des (codes de) formules ou d'énoncés. Et il est loin d'être clair à nos yeux qu'on puisse toutes les remplacer par des schémas d'axiomes et faire l'économie d'un prédicat décitationnel permettant l'ascension sémantique et l'élimination des guillemets (ou plutôt en l'occurrence le passage des codes à la Gödel $\ulcorner\varphi\urcorner$ à l'énoncé φ lui-même). Mais peut-être qu'une telle reformulation « déflationnistiquement acceptable » de l'argument inductif de Shapiro et Ketland est possible. Notez toutefois, que de toute façon dans PA^{+ind}_{Tar}, on ne se contente pas de *postuler* que tous les axiomes de T sont vrais et que les règles préservent la vérité, on le *prouve* (voyez à nouveau Halbach (2014, chapitre 8) pour plus de détails sur la manière dont on dérive ces énoncés à partir des axiomes aléthiques compostionnels).

s'appuyer sur la notion d'acceptation d'une théorie et sur les divers engagements que nous prenons lorsque nous adoptons une théorie de base T. Après avoir reconnu qu'une extension aléthique réflexive à la Tarski donne bien une justification possible des schémas de réflexion, Tennant insiste sur le fait que ce type de justification n'est pas la seule possible, et que d'autres routes sont ouvertes au déflationniste :

> [...] le déflationniste préfère la voie modeste. Il réfléchit simplement sur ses méthodes de preuves actuellement cantonnées à $[T]$, et apprécie ce qu'il voit. Il a confiance en elles. Il peut exprimer cette confiance au moyen de divers principes *mathématiques* nouveaux, de forces logiques variables, et qui tous évitent tout emploi d'un prédicat de vérité. [...] parmi ceux-ci se trouve le *principe de correction* suivant :
>
> $$Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$$
>
> Le déflationniste pourrait fort bien souhaiter adopter toutes les instances de ce schéma. Après tout, il était disposé à asserter n'importe quel énoncé φ pour lequel il avait fourni une preuve dans T, pourquoi ne pas alors être également disposé à asserter tout énoncé φ pour lequel il peut fournir une preuve du fait que l'énoncé φ est susceptible de recevoir une preuve dans $[T]$? (Tennant, 2005, p. 91)
>
> [...] Rien à voir avec une notion substantielle de vérité ici. Tout ce que l'on fait ici, c'est appliquer le raisonnement réflexif[209] à l'œuvre dans le principe de correction. [...] Il n'y a qu'une réflexion sur nos engagements axiomatiques et déductif actuels, et une tentative d'articuler la conséquence systématique de ces engagements, en composant le principe de réflexion apparemment plutôt faible $[\mathrm{RFN}_{\mathrm{P.R}}(\mathrm{T})]$. Le prédicat de vérité ne fait même pas ne serait-ce qu'une apparition dans ces réflexions. [...] Aucune justification supplémentaire n'est nécessaire pour le nouvel engagement pris en exprimant nos précédents engagements. Dès l'instant qu'on apprécie à sa juste valeur le processus de réflexion, et la façon dont son aboutissement est exprimé par le principe de réflexion, on possède déjà une explication des raisons pour lesquelles quelqu'un qui accepte $[T]$ devrait également accepter toutes les instances du principe de réflexion.

209. Littéralement : « *reflective thought* ».

> [...] Plus nous essayons d'exprimer notre confiance dans nos méthodes licites de preuve, plus nous les étendons. Le concept même de 'méthode licite de preuve' est indéfiniment extensible. Nous n'avons pas besoin d'un concept substantiel de vérité pour établir ce résultat ; et ce résultat s'applique à toutes les méthodes de preuve, y compris celle qui évitent tout rapport explicite avec la vérité. (Tennant, 2005, p. 92)

Ces passages extraits de Tennant (2005, p. 91 & 92) et de Tennant (2002, p. 575) regroupent l'ensemble des indications (quelque peu succinctes selon nous) fournies par Tennant sur la façon dont on pourrait envisager de justifier les principes de réflexion sans faire appel à la notion de vérité. Pour ce faire, Tennant propose de s'appuyer sur ce qu'il appelle un « raisonnement réflexif » ou « processus de réflexion ». Ce processus est visiblement censé s'appliquer directement à nos méthodes de preuve, indépendamment de toute notion de vérité. Il consiste apparemment à examiner certains « engagements » que nous prenons lorsque nous acceptons, ou adoptons, ou peut-être faudrait-il dire lorsque nous plaçons notre confiance en une théorie de base. Selon les indices laissés par Tennant, la conséquence systématique de ces engagements débouche sur la formulation et l'adoption de principes de réflexion, dans lesquels le prédicat de vérité ne prend aucune part. Une fois correctement comprise la nature de ce processus de réflexion exercé sans mettre en jeu la notion de vérité, nous sommes aussitôt en possession d'une justification, ou d'une explication des raisons pour lesquelles nous devons étendre notre théorie de base au moyen d'un principe de réflexion.

Autant le dire tout de suite : si les remarques de Tennant nous paraissent extrêmement intéressantes, stimulantes et suggestives, et si elles peuvent nous sembler tout à fait plausibles à première vue, elles sont cependant loin de suffire à nos yeux pour lever les doutes sur la manière dont les principes de réflexion peuvent être justifiés. Quelle est en effet la nature de ce processus de réflexion sur lequel Tennant entend s'appuyer ? Sommes-nous sûrs qu'il peut s'exercer sans faire intervenir la notion de vérité. [210] De même, on peut se demander sur quelles théories ce processus peut s'exercer : sur n'im-

210. *A contrario*, Ketland ou un partisans d'une théorie aléthique substantielle réflexive pourrait soutenir que le processus en question consiste précisément à prendre conscience que les axiomes de T sont vrais et que les règles d'inférence de T préservent la vérité pour en *déduire* que tous les théorèmes de T sont vrais et établir divers principes de réflexion pour T. Ne conserver que l'appellation de ce processus ainsi que le résultat final sous la forme de schémas exprimés dans $\mathcal{L}_T$ apparaît comme une manière assez peu *fair play* et assez peu convaincante de se débarasser de la notion de vérité. Il faut donc, nous semble-t-il, fournir plus d'explication sur la nature de ce processus de réflexion et sur les ressources qu'il mobilise.

porte quelle théorie que nous pourrions accepter ? Mais accepter en quel sens ? Au sens où nous l'aurions choisie arbitrairement ? Au sens où nous pourrions légitimement —mais alors, à quelles conditions ?— l'adopter comme une théorie scientifique satisfaisante ? La fin de l'extrait cité ci-dessus suggère que le processus de réflexion est censé s'appliquer à nos méthodes *licites* de preuve.[211] Mais qu'est-ce qu'une méthode de preuve licite ? Une réponse classique et même peut-être assez naturelle, consisterait à dire que sont « licites » les méthodes de preuve qui conduisent à la vérité, ou plus spécifiquement que sont « licites » les méthodes de preuve qui mènent de prémisses vraies à des conclusions vraies. Si on adopte une telle caractérisation, il semble néanmoins qu'on soit ramené à la case départ dans la tentative de réhabilitation du déflationnisme proposée par Tennant.

Parallèlement, quelle est la nature de ces engagements que nous sommes supposés avoir contractés lorsque nous « acceptons » une théorie ? Comment et dans quelle mesure notre attitude épistémique vis-à-vis d'une théorie T, que ce soit croyance, défiance, confiance, acceptation, refus, adoption pour les besoins d'un argument, supposition raisonnable, etc... peut-elle avoir un impact sur la correction de ladite théorie (si tant est que ce soit bien la correction de T que les principes de réflexion sont censés exprimer) ? On le voit, la démarche présentée par Tennant soulève de nombreuses questions. Et ce n'est pas lui faire un mauvais procès que de considérer que des clarifications supplémentaires sont exigibles.

S'il faut donc dresser un bilan des échanges opposant Ketland et Tennant, nous sommes tentés, au moins provisoirement, de les renvoyer dos à dos : contre Ketland, nous sommes d'accord avec Tennant pour dire que le simple fait qu'une justification des principes de réflexion soit disponible dans une extension aléthique assez forte ne suffit pas à montrer que c'est la seule possible et que la vérité est indispensable à mener une justification de ces principes. En ce sens, la stratégie suggérée par Tennant, si on l'interprète comme une tentative de contourner l'argument sémantique[212] en montrant qu'une justification alternative des principes de réflexion est possible, nous semble parfaitement envisageable, prometteuse pour le déflationniste, et digne d'être explorée. En revanche, contre Tennant, nous sommes d'accord avec Ketland pour considérer que le projet annoncé est loin d'avoir été mené à bien pour le moment. Les remarques formulées par Tennant (2002, 2005) sont trop succinctes en l'état pour nous avoir convaincus qu'une

211. Et par cette application donner l'occasion d'étendre l'extension du concept de méthodes de preuve licites lui-même.
212. et non pas sémantique*.

justification des principes de réflexion acceptable en termes purement déflationniste est accessible. À la fin des échanges entre Ketland (2010) et Tennant (2010), la question de savoir si les principes de réflexion peuvent être justifiés sans faire appel à une notion substantielle de vérité reste donc largement ouverte à nos yeux. Nous allons voir que ce fut justement une voie de recherche fertile donnant lieu à d'interessants développements : plusieurs contributions récentes au débat sur le deflationnimse peuvent en effet être considérées comme des tentatives de compléter la démarche de Tennant face aux critiques de Ketland.

4.3.3 Développements

Commençons par revenir sur la formulation précise de la stratégie développée par Tennant. Au vu de ce qui précède, il nous paraît indéniable que la clef de l'argumentation de Tennant est la mise au jour d'un processus de réflexion ne faisant aucun usage de la notion de vérité. Pour autant, Tennant est étrangement peu disert et ne donne guère de détails sur la nature et le fonctionnement d'un tel processus. Au point qu'on a parfois le sentiment que c'est là une pièce manquante de l'édifice qu'il tente de construire. Malgré tout, Tennant donne quelques indications, pour la plupart concentrées dans les pages que nous avons déjà citées ci-dessus. Le passage crucial, le seul où Tennant dépeint les étapes de ce processus, nous semble être celui extrait de Tennant (2005) où il décrit un agent ayant adopté une théorie de base T comme s'étant placé dans une position où il est disposé à asserter n'importe quel énoncé φ pour lequel il peut fournir une preuve dans T. Dès lors, interroge Tennant,

> [P]ourquoi ne pas alors être également disposé à asserter tout énoncé φ pour lequel il peut fournir une preuve du fait que l'énoncé φ est susceptible de recevoir une preuve dans $[T]$? (Tennant, 2005, p. 91)

Cette dernière question est assurément purement rhétorique et il est clair que la réponse ne fait aucun doute aux yeux de Tennant : quiconque ayant accepté tous les théorèmes de T devrait également accepter les énoncés φ dont il peut prouver qu'ils sont des théorèmes de T. On retrouve ici sous une autre forme la notion d'obligation épistémique conditionnelle formulée par Ketland.[213] Cette idée qu'il existe des engagements implicites que nous contractons lorsque nous adoptons une théorie semble à vrai dire bénéficier d'un certain consensus au sein de la communauté des philosophes discutant les thèses

213. Et inspirée, nous l'avons déjà signalé, des travaux de Feferman.

déflationnistes à la lumière des phénomènes d'incomplétude.[214] Mais si chez Ketland l'analyse de ces engagements implicites s'appuyait sur la notion de vérité,[215] Tennant entend évidemment éviter ce détour par une extension aléthique. Le passage proposé ici de l'acceptation de tous les théorèmes de T à l'acceptation de tous les énoncés dont on peut prouver qu'ils sont des théorèmes de T peut sembler plausible. On peut toutefois s'interroger sur la formulation précise du processus que Tennant veut décrire.

Admettons qu'un agent X ait « accepté » une théorie de base T —disons $T = PA$ pour fixer les idées. Admettons que cette notion d'acceptation se traduise ou se manifeste par le fait que X est disposé à asserter tout théorème de PA, ce qu'on pourrait décrire par

(i) $PA \vdash \varphi \Rightarrow X$ est disposé à asserter φ.

Admettons également que de ceci, il découle de manière relativement directe que X devrait être disposé à asserter tout énoncé φ dont il peut prouver que c'est un théorème de PA, ce qu'on pourrait décrire par

(ii) $PA \vdash \ Thm_{PA}(\ulcorner\varphi\urcorner) \Rightarrow X$ est disposé à asserter φ.[216]

214. Ce qui ne veut pas dire qu'elle soit unanimement acceptée : pour une critique de l'idée que de tels engagements implicites s'imposent à nous quelle que soit par ailleurs notre conception des mathématiques, voyez Dean (2015) qui défend l'idée que pour un tenant du finitisme strict à la Tait ou un partisan des thèses d'Isaacson sur la complétude épistémique de l'arithmétique de Peano du premier ordre, de telles obligations n'ont pas lieu d'être.

215. Ce qu'on pourrait peut-être résumer comme ceci :

> Si j'adopte un théorie T, *au sens où je la tiens pour vraie*, alors un certain nombre de conséquences, non directement dérivables dans T, s'imposent à moi.

216. Au demeurant, on peut se demander ici comment il faut interpréter précisément la phrase « tout énoncé φ dont X peut prouver que c'est un théorème de PA » : prouver à partir de quelles ressources exactement ? Supposons que X ait accepté PA, à la manière de l'étape (i) ci-dessus, quel cadre peut-il employer pour prouver $Thm_{PA}(\ulcorner\varphi\urcorner)$? Autrement dit, comment formuler

$$\text{(ii)}\ \mathbf{??} \vdash \ Thm_{PA}(\ulcorner\varphi\urcorner) \Rightarrow X \text{ est disposé à asserter } \varphi\,?$$

Si X a accepté PA, sans doute peut-il employer cette théorie pour établir des énoncés du type $Thm_{PA}(\ulcorner\varphi\urcorner)$, *i.e.* pour établir que φ est prouvable dans PA. Mais peut-être X peut-il se prévaloir d'autres ressources. Peut-être devrait-il au contraire se cantonner à des ressources plus limitées. Ce pourrait être le cas par exemple si X avait accepté à l'étape (i) une théorie $\mathcal{T}$ très « risquée », alors qu'il voudrait rester plus prudent et employer à l'étape (ii) une théorie plus « sûre » pour raisonner sur la prouvabilité dans $\mathcal{T}$. Nous laissons cette question en suspens.

Signalons toutefois que, sous certaines hypothèses raisonnables concernant la manière dont la syntaxe de $\mathcal{L}_{PA}$ est codée et la façon dont $Thm_{PA}(x)$, le prédicat de prouvabilité dans PA, est défini, un énoncé du type $Thm_{PA}(\ulcorner\varphi\urcorner)$ sera généralement beaucoup plus facile à démontrer que l'énoncé φ lui-même. Ce n'est au fond pas très étonnant puisqu'une fois le codage fixé, quelle que soit la complexité de φ, l'énoncé $Thm_{PA}(\ulcorner\varphi\urcorner)$ est toujours quant à lui un énoncé Σ^0_1. On peut d'ailleurs donner un sens précis à tout ceci en termes de longueur ou de complexité des preuves. Sur ce point particulier, voir par exemple Parikh (1971, Théorème 1.3, p. 496).

Quel lien y a-t-il entre ceci et l'acceptation par X de toutes les instances d'un schéma d'axiomes du type

$$Thm_{PA}(\ulcorner \varphi \urcorner) \rightarrow \varphi \, ?$$

Si elle peut paraître plausible, nous allons voir que la connection est loin d'être aussi limpide qu'il peut sembler.

4.3.3.1 La (re)formulation de Cieśliński (2010)

En s'appuyant sur ces éléments esquissés par Tennant (2002, 2005), Cieśliński (2010) a proposé d'interpréter plus précisément l'argumentation de Tennant de la manière suivante :

> La proposition de Tennant contient deux éléments — l'un descriptif et l'autre normatif. Sur le plan descriptif, le processus démarre par une réflexion sur mes engagements déductifs en tant qu'utilisateur de PA : je suis prêt à accepter tout énoncé φ pour lequel je peux fournir une preuve dans PA. Formulons ceci explicitement. À la première étape du processus de réflexion j'accepte la déclaration suivante :
>
> (D) Pour tout énoncé φ, si φ possède une preuve dans PA, alors je suis prêt à accepter φ.
>
> Tel que je le conçois, le statut de (D) est descriptif. C'est une affirmation factuelle, portant sur la manière dont j'emploie les axiomes de PA et ses moyens de preuves. Il se peut que je parvienne à (D) par introspection ou par une sorte de généralisation empirique —cela n'a pas d'importance. Dans ce qui suit je supposerai simplement que je peux effectivement arriver à (D) sans employer aucun concept de vérité (mais uniquement le concept pragmatique d' « accepter » ou d' « asserter » un énoncé). On pourrait dire en fait que (D) exprime simplement ma confiance en PA et ses moyens de preuve.
>
> À l'étape suivante du processus arrive la formalisation : je réalise que le contenu de (D) (ou une partie de ce contenu) peut être exprimé(e) par l'ensemble infini des énoncés arithmétiques de la forme « $Thm_{PA}(\ulcorner \varphi \urcorner) \rightarrow \varphi$ »—appelons-le l'ensemble des axiomes réflexifs. La thèse de la formalisation est :

(F) L'ensemble des axiomes réflexifs exprime le contenu de (D) (ou une partie de celui-ci) .

À présent vient la thèse normative :

(P) Quiconque accepte PA devrait également accepter toutes les instances du schéma de réflexion.

L'argument pour (P) est le suivant : nous remarquons que toute personne qui accepte PA devrait également accepter (D). La raison est que (D) exprime simplement le fait que la personne en question accepte PA ; et l'affirmation serait que les données sur lesquelles (D) s'appuie, qu'elles soient introspectives ou empiriques, sont en principe facilement accessibles à n'importe quel être humain rationnel, ce serait donc une grave erreur de les ignorer. En fait, puisque j'ai une raison d'accepter (D), alors d'après (F) j'ai également une raison d'accepter tous les axiomes réflexifs. (Cieśliński, 2010, Section 3.2 p. 416-417)

Par son niveau de détails plus fouillé, l'élaboration proposée par Cieśliński se prête plus facilement à une analyse critique. Si nous l'avons bien compris l'argument a la forme suivante :

Argument de Tennant-Cieśliński :

(1) J'accepte PA
(2) Je dois accepter :
(D) Si φ est un théorème de PA alors j'accepte φ
(3)/(F) Rfn(PA) exprime le contenu de (D)
(4) Je dois accepter Rfn(PA)

(CL) Si j'accepte PA, je dois accepter Rfn(PA)

Bien évidemment, ce squelette argumentatif ne constitue pas *à soi seul* une déduction en bonne et due forme. Le passage de (1) à (2) par exemple, n'est pas une inférence déductive et requiert, semble-t-il, des éclaircissements supplémentaires. En outre, la prémisse (3) —notée (F) pour « formalisation » dans Cieśliński (2010)— n'est pas si évidente qu'on pourrait la prendre comme une sorte d'axiome sans fournir à son appui quelque explication ou justification complémentaires. Cieśliński lui-même le reconnaî puisqu'il déclare à la suite du passage que nous venons de citer :

> Pour évaluer l'argument ci-dessus, *la question cruciale est : qu'entendons-nous ici par « accepter PA » ?* L'interprétation naturelle se déploie comme suit : accepter PA signifie être prêt à accepter tout énoncé pour lequel une preuve à partir des axiomes de PA peut être fournie. Dans cette approche, c'est (D) qui nous donne la signification de « j'accepte PA ». (Cieśliński, 2010, Section 3.2 p. 417, nous soulignons)

Ceci pour justifier le passage de (1) à (2). Concernant la prémisse (3)/(F), Cieśliński déclare que si on interpréte « j'accepte PA » par (D), l'argumentation de Tennant lui paraît convaincante, mais il ajoute dans une note de bas de page :

> *Ceci sous réserve que nous prenions (F) pour acquis.* En effet, on pourrait encore se demander en quel sens exactement les axiomes réflexifs expriment une partie du contenu de (D) : que signifie « exprimer » ici ? *C'est une question complexe que je ne discuterai pas dans cet article* —je me concentrerai uniquement sur la tâche consistant à montrer ce qu'on peut réaliser *si on accepte (F) comme donné.* (Cieśliński, 2010, p. 417, note 6, nous soulignons)

En d'autres termes, Cieśliński reconnaît lui aussi le caractère crucial de la notion d'acceptation dans le passage de (1) à (2) et dans la justification de la prémisse (3), tout comme l'importance fondamentale de la signification du terme « exprimer » dans ce raisonnement. Mais il ne donne finalement guère de détails lui non plus : le passage de (1) à (2) est justifié au moyen de ce qui ressemble fort à une pure et simple stipulation (accepter PA c'est accepter (D)...), tandis qu'une justification détaillée du fait que Rfn(PA) « exprime » le contenu de (D) est renvoyée à une autre occasion. [217]

217. À la décharge de Cieśliński (2010), soulignons que cette reconstruction ou reformulation rapide de l'argument originel de Tennant (2002, 2005) ne constitue pas l'unique ni même le principal objectif de son article. Le texte contient d'autres réflexions intéressantes ainsi que des résultats techniques nouveaux apportant un autre éclairage sur la stratégie de Tennant et sur ses limites. En voici un résumé des principaux points.

1. Tout d'abord, Cieśliński (2010) montre que non seulement une théorie conservative sur l'arithmétique de Peano ne peut pas permettre de montrer que tous les théorèmes de PA sont vrais, mais encore qu'une telle théorie conservative ne peut même pas établir que tous les théorèmes (exprimés dans $\mathcal{L}_{PA}$) de la logique (*i.e.* prouvables à partir d'un ensemble vide de prémisses) sont vrais. Autrement dit, une théorie conservative sur PA ne peut établir « tous les théorèmes de T sont vrais » pour *aucune théorie T exprimée dans $\mathcal{L}_{PA}$*. Ce résultat prend précisément la forme du théorème suivant :

   Théorème. (Cieśliński, 2010, p. 412)

   $$PA_{Tar} + \forall x[En(x) \wedge Thm_{\emptyset}(x) \rightarrow Vr(x)] \vdash Ref(PA)$$

   où « $\forall x[En(x) \wedge Thm_{\emptyset}(x) \rightarrow Vr(x)]$ » est un énoncé de $\mathcal{L}_{PA} \cup \{Vr\}$ formalisant l'assertion « tous

Face à cette réinterprétation de la plaidoirie de Tennant, Ketland (2010) conteste à la fois le passage de (1) à (2) et la prémisse (3)/(F). Concernant le passage de (1) à (2), remarquons qu'une fois formalisée, une théorie se présente généralement sous la

les théorèmes de la théorie vide (*i.e.* de la logique) exprimée dans $\mathcal{L}_{PA}$ sont vrais » ; « $Thm_{\emptyset}(x)$ » est donc une formule de $\mathcal{L}_{PA}$ représentant la propritété « x est un énoncé de $\mathcal{L}_{PA}$ prouvable à partir de l'ensemble vide de prémisses ». Comme précédemment, $Ref(PA)$ dénote ici le principe de réflexion global pour l'arithmétique de Peano.

Reprenant un résultat de Kotlarski (1986) Cieśliński (2010, p. 414) précise ensuite la force logique de l'extension obtenue à partir de PA_{Tar} en lui ajoutant le principe de réflexion global et donne diverses formulations de cette théorie toutes équivalente à Δ_0-PA_{Tar}, *i.e.* à la théorie PA_{Tar} à laquelle on a ajouté l'induction pour les formules du langage étendu (contenant éventuellement le prédicat de vérité) mais uniquement de complexité Δ_0 —Autrement dit, Δ_0-$PA_{Tar} := PA_{Tar}+ SI(\mathcal{L}_{Vr} \cap \Delta_0)$ est la théorie PA_{Tar} augmentée de l'induction pour les formules de $\mathcal{L}_{Vr}$ ne contenant que des quantifications bornées. C'est bien entendu une théorie strictement plus faible que PA^{+ind}_{Tar}.

2. Après avoir critiqué l'argument de Tennant (2002, 2005) dans sa forme originale en soulignant que la construction proposée par Tennant ne rend pas justice au rôle (expressif) indispensable que les déflationnistes attribuent au prédicat de vérité et qu'elle brise la distinction entre, d'une part, déflationnisme et, d'autre part, éliminativisme ou théorie de la redondance en matière de vérité, Cieśliński (2010) propose une nouvelle variante de la stratégie de Tennant (2002, 2005) également inspirée de Field (1999) : il s'agit d'employer un schéma de réflexion pour la logique (dans le langage aléthique $\mathcal{L}(Vr)$) pour justifier l'extension de la théorie PA_{Tar} (conservative sur PA) à la théorie plus forte PA^{+ind}_{Tar}. Cette stratégie s'appuie sur le résultat technique suivant :

   Observation. *(Cieśliński, 2010, p. 419) Soit $Ref_{\mathcal{L}(Vr)}(Log)$ l'ensemble de toutes les instances du schéma*

   $$Ref_{\mathcal{L}(Vr)}(Log) : \forall x[Thm_{\emptyset}(\ulcorner \varphi(\dot{x}) \urcorner) \rightarrow \varphi(x)]$$

   où $\varphi(x)$ est une formule à une variable libre du langage étendu $\mathcal{L}(Vr) := \mathcal{L}_{PA} \cup \{Vr\}$, *i.e. du langage contenant le prédicat de vérité Vr.*
   Alors

   $$PA_{Tar}+Ref_{\mathcal{L}(Vr)}(Log)$$

   prouve toutes les instances de l'induction pour le langage étendu $\mathcal{L}(Vr)$. D'où

   $$PA^{+ind}_{Tar} \subseteq PA_{Tar}+Ref_{\mathcal{L}(Vr)}(Log).$$

   Notez que $Ref_{\mathcal{L}(Vr)}(Log)$ n'est autre qu'un schéma de réflexion uniforme sur la théorie vide pour le *langage étendu* $\mathcal{L}(Vr)$:

   $$Ref_{\mathcal{L}(Vr)}(Log) = \mathrm{RFN}_{\mathcal{L}(Vr)}(\emptyset).$$

   Cieśliński (2010, p. 419) contient une preuve de cette observation. On peut néanmoins la voir comme un cas particulier d'un résultat plus général dû à Kreisel et Lévy (1968, p. 105-106).

Nous laisserons toutefois ces aspects du travail de Cieśliński (2010) de côté, car ce qui nous intéresse ici est véritablement le problème de la justification des schémas de réflexion (principalement sur une théorie arithmétique, mais sans exclusive : la question se pose tout aussi bien à nos yeux pour une théorie logique ou autre, dès lors qu'on dispose *a minima* des moyens de coder la syntaxe du langage et le système de preuve de la théorie en question). Dans cette optique, la point 2. ci-dessus de Cieśliński (2010) ne fait que déplacer le problème de la justification des schémas de réflexion sur PA —sans faire usage d'une notion substantielle de vérité— à celui de la justification —sans faire usage d'une notion substantielle de vérité— de ces mêmes schémas pour la logique. Tennant (2010) semble d'ailleurs en plein accord avec nous sur ce point.

forme d'un ensemble d'axiomes (ensemble fini voire récursif, ou à tout le moins récursivement énumérable) accompagné de règles de dérivation (logiques et autres) permettant de « produire » une infinité de théorèmes. Que signifie « accepter » ce type d'objet linguistico-syntaxique ? Si on « accepte » les axiomes et les règles d'inférence d'une théorie, pouvons-nous ou devons-nous également être considérés comme « acceptant » toutes ses conséquences déductives aussi lointaines soient-elles ? À supposer que la conjecture de Goldbach soit un théorème de PA, puis-je être dit avoir accepté —et en quel sens ?— cet énoncé arithmétique dès lors que j'ai accepté PA, alors même qu'aucune preuve n'en est à ce jour connue ? Ketland (2010, p. 429) soutient qu'en l'absence d'une caractérisation plus poussée de la notion d'acceptation, l'hypothèse de clôture déductive portant sur cette notion est loin d'aller de soi. Il faut évidemment supposer ici que certaines hypothèses normatives sur la notion d'acceptation sont à l'œuvre. Et il faut garder à l'esprit que l'on parle certainement ici d'une notion d'acceptation en droit, appuyée sur une dose non négligeable d'idéalisation, et qu'on ne cherche pas à donner une quelconque description réaliste de la psychologie réelle du mathématicien ou de l'homme de sciences. Peut-être conviendrait-il de parler d'acceptation rationnelle, ou d'acceptation légitime, ou d'acceptation en principe... Toutefois, nous ne voulons pas nous étendre plus que nécessaire sur ce point précis : il nous semble en effet plausible qu'il existe un sens raisonnable du terme « accepter » d'après lequel accepter une théorie c'est accepter ou s'engager à accepter (au moins) tous ses théorèmes. Ceci nous semble correspondre à la pratique, certes idéalisée, des mathématiciens et d'ailleurs il est tout à fait courant en logique d'identifier une théorie, *qua* axiomes + règles d'inférence, avec l'ensemble des ses théorèmes ou même de définir directement une théorie comme un ensemble déductivement clos d'énoncés. Supposons donc qu'accepter une théorie ce soit (au moins) accepter (peut-être implicitement) tous ses théorèmes. Acceptons également que (D) ne soit que l'énonciation de ceci, et qu'une fois PA acceptée nous soyons donc tenus d'accepter (D). Autrement dit, tenons le passage de (1) à (2) pour non problématique.

Ce qui peut paraître beaucoup plus discutable en revanche, c'est l'inférence de (2) à la conclusion (4) modulo la prémisse (3)/(F). Comme nous l'avons déjà dit, Cieśliński (2010, p. 417, note 6) reconnaît lui aussi l'importance des difficultés soulevées par l'hypothèse (3)/(F) mais renonce, dans le cadre de son article, à apporter plus d'éclaircissements ou d'explications à son sujet. Ce n'est que dans Cieśliński (2017) qu'il reviendra sur ce problème et proposera sa solution entièrement développée au problème de la

conservativité.

4.3.3.2 Variations sur le thème de l'acceptation

Avant même d'en venir à une analyse plus précise de l'argument de Tennant-Cieśliński, en particulier de la prémisse (3)/(F), et d'examiner ce qui a pu être dit à son sujet par les protagonistes de notre débat, remarquons déjà qu'il est en tout cas possible de proposer un modèle plausible de l'acceptation qui *ne* valide *ni* la prémisse (3), *ni* la conclusion (4).

C'est ce qu'on pourrait appeler une modélisation *procédurale* de la notion d'acceptation. Supposons qu'un agent cognitif X est muni d'une « *belief*[218] *box* », *i.e.* d'une boîte à croyances, et que cet agent puisse être dit accepter (le contenu d') un énoncé φ si et seulement si φ est placé dans sa boîte à croyances. À présent supposons que l'agent X accepte une théorie T au sens précis suivant : au départ, cet agent place l'ensemble des axiomes de T dans sa boîte. Mais dans le même temps, X accepte les règles de déduction de T. Une règle ne se présente pas sous la forme d'un énoncé. Accepter une règle, ce n'est donc pas tout à fait la même chose qu'accepter un énoncé. Néanmoins, pourrait-on dire, accepter une règle, c'est accepter de la suivre, en l'occurrence c'est s'engager à accepter tout énoncé obtenu par application de la règle.[219] Au cours du temps donc, X partant des axiomes de T applique les règles de déduction de cette théorie et prouve des théorèmes qu'il intègre aussitôt à sa boîte à croyances. Si l'on fait abstraction des limitations de ressources cognitives, du temps passé, de la complexité des preuves, *etc.* l'output linguistique total accepté par X, le contenu final de sa boîte à croyances, sera exactement l'ensemble des théorèmes de T *et rien d'autre*, ce qu'on nomme habituellement la clôture déductive de T, noté $T^{\vdash} = \{\varphi \in \mathcal{L}_T \mid T \vdash \varphi\}$.

Toutefois, pourrait-on dire, le mécanisme que nous venons de décrire et qui mène à $T^{\vdash}$ ne représente, peut-être, que la première étape de l'acceptation. Il ne prend pas encore en compte le processus de réflexion à l'œuvre dans le passage de PA à $\mathrm{Rfn}(PA)$ (ou plus généralement de T à $\mathrm{Rfn}(T)$). Il est vrai que si l'on suppose que sont codables[220] dans le langage $\mathcal{L}_T$ lui-même, la syntaxe de $\mathcal{L}_T$ et les règles de déductions suivies par X qui

218. Ou peut-être d'une « *acceptation box* », si l'on veut maintenir une distinction entre la croyance et l'acceptation...

219. D'où notre appellation de modélisation *procédurale* de la notion d'acceptation d'une théorie.

220. Voyez le chapitre précédent pour plus de détails sur sur ce que doit contenir T comme « assez d'arithmétique » pour pouvoir coder sa propre syntaxe, et notamment sur les hypothèses de représentabilité des relations récursives dans T.

président à l'inclusion dans sa boîte à croyances, *i.e.* les règles de dérivation de T, on peut alors définir un prédicat de prouvabilité pour T. On peut même le construire dans $\mathcal{L}_T$ de manière suffisamment raisonnable (ou canonique pour reprendre la terminologie en vigueur en cette matière) pour que ce prédicat $Thm_T(x)$ satisfasse le critère de Kreisel (1953, p. 405) portant sur l'expression de la prouvabilité dans une théorie T. Ce critère consiste à exiger que pour tout énoncé φ de $\mathcal{L}_T$,

$$K1 \quad T \vdash Thm_T(\ulcorner\varphi\urcorner) \text{ ssi } T \vdash \varphi \text{ [221]}$$

À l'issue de la première étape ci-dessus de notre modèle de l'acceptation, la boîte à croyance contient pour l'instant exactement les théorèmes de T. Si on note $\mathcal{B}$ le contenu de cette boîte, on a donc

$$\varphi \in \mathcal{B} \text{ [222] ssi } T \vdash \varphi$$

Associé à $K1$ ci-dessus, il suit immédiatement :

$$(K1') \quad \varphi \in \mathcal{B} \text{ ssi } T \vdash \varphi \text{ ssi } T \vdash Thm_T(\ulcorner\varphi\urcorner)$$

À présent, remarquons qu'avec *ce* modèle, simple mais plausible de la notion d'acceptation, lorsque Tennant (et Cieśliński) écrivent que si X a accepté tous les théorèmes de

221. Remarquez qu'il vaudrait mieux que Thm_T satisfasse ce critère. Dans le cas contraire, il ne serait guère satisfaisant comme prédicat de prouvabilité. En effet, si

$$T \vdash Thm_T(\ulcorner\varphi\urcorner) \text{ mais } T \nvdash \varphi$$

cela signifie qu'il existe un énoncé non prouvable dans T tel que T « croit » néanmoins à tort que c'est un théorème. Et si

$$T \nvdash Thm_T(\ulcorner\varphi\urcorner) \text{ alors que } T \vdash \varphi$$

c'est que, bien que T puisse prouver φ, T « manque » de voir que φ est un de ses théorèmes. Notons néanmoins que si $Thm_T(x)$ est défini par $Thm_T(x) := \exists y Dem_T(y, x)$ à partir d'un prédicat $Dem_T(x, y)$ représentant la relation de démonstration dans T au sens où pour toute paire d'entiers m, n,

$$T \vdash Dem_T(\overline{m}, \overline{n}) \text{ ssi } m \text{ est le code d'une preuve dans } T \text{ de la formule codée par } n,$$

autrement dit si Thm_T est raisonnablement construit comme nous l'avons fait jusqu'à présent, il est alors facile de montrer que $Thm_T(x)$ satisfait le critère de Kreisel.

222. Autrement dit φ est accepté par X.

T, il devrait accepter tous les énoncés φ dont il *peut prouver qu'ils sont des théorèmes de* T [223], ils n'ont fourni strictement *aucune* justification pour l'adoption d'un principe de réflexion tel que Rfn(T).

En effet, si nous interprétons « X a accepté tous les théorèmes de T » (autrement dit l'étape 1 de notre modèle de l'acceptation) par

1. X a placé dans sa boîte à croyance tous les théorèmes de T

et que nous interprétons « X devrait accepter tous les énoncés pour lesquels il peut fournir une preuve du fait qu'ils sont prouvables dans T » (c'est-à-dire l'étape deux telle qu'elle est suggérée pas Tennant et Cieśliński) par

2. X place (ou doit placer) dans sa boîte à croyance tous les énoncé φ tels qu'il peut prouver $Thm_T(\ulcorner\varphi\urcorner)$, c'est-à-dire tous les énoncés dont il peut prouver qu'ils sont des théorèmes de T.

alors l'équivalence $(K1')$ ci-dessus montre qu'on n'aura, ce faisant, pas progressé d'un pouce. Si le « processus de réflexion » consiste à accepter tous les énoncés dont on peut montrer qu'ils sont des théorèmes de T, autrement dit si X accepte « à la réflexion » tous les énoncés φ tels qu'il peut prouver $Thm_T(\ulcorner\varphi\urcorner)$, il ne fera qu'ajouter au contenu de sa boîte des énoncés qui y sont déjà présents.

Au demeurant, ce type de traitement de la réflexion au moyen d'une *règle procédurale* consistant à accepter/asserter φ dès lors qu'on a établi $Thm_T(\ulcorner\ \varphi\urcorner)$ —c'est-à-dire dès lors qu'on a *prouvé* (dans T) $Thm_T(\ulcorner\ \varphi\urcorner)$— est parfois désigné dans la littérature sur les

223. Redonnons le passage précis de Tennant (2005) auquel nous faisons allusion ici :

> [...] le *principe de correction* suivant :
>
> $$Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$$
>
> Le déflationniste pourrait fort bien souhaiter adopter toutes les instances de ce schéma. Après tout, il était disposé à asserter n'importe quel énoncé φ pour lequel il avait fourni une preuve dans T, pourquoi ne pas alors être également disposé à *asserter tout énoncé φ pour lequel il peut fournir une preuve du fait que l'énoncé φ est susceptible de recevoir une preuve dans [T]* ? (Tennant, 2005, p. 91)

On trouve exactement le même type d'argumentation dans Cieśliński (2010) :

> L'idée est la suivante : commencez avec la théorie S que vous utilisez actuellement ; puis réfléchissez sur vos engagements déductifs et axiomatiques et tentez de les exprimer sous la forme d'un principe de réflexion approprié. Au cours de ce processus de réflexion, vous remarquez que vous êtes prêt à accepter tout énoncé φ pour lequel vous pouvez produire une preuve dans S. Ceci vous donne une raison d'*accepter tout énoncé φ pour lequel vous pouvez montrer qu'il est possible de produire une preuve dans S*. Ce faisant, vous parvenez à une théorie S^* qui est une approximation raisonnable de l'affirmation « Tous les théorèmes de S sont vrais ». (Cieśliński, 2010, p. 411)

schémas de réflexion sous le nom de *règle de Parikh*, ou *règle* de réflexion *locale* .[224] Ce n'est donc pas une conception totalement farfelue ou *ad hoc* de la notion d'acceptation et de la façon dont on peut la modéliser. Néanmoins, ce genre de réflexion donnée sous forme de *règle* est bien connu pour être plus faible que les schémas de réflexion locaux ou uniformes, et pour produire des *extensions conservatives* dès lors que T contient un minimum d'arithmétique. La faiblesse de ce type de *règle* de réflexion a eu pour conséquence qu'elle a beaucoup moins excité la curiosité des logiciens et philosophes que les schémas de réflexion du type $\mathrm{Rfn}(T), \mathrm{RFN}(T)$ ou $Ref(T)$. Malgré cela, l'extension d'une théorie de base au moyen de *règle* de réflexion de cette sorte peut avoir un intérêt mathématique en dépit de sa conservativité dans la mesure où elle permet généralement un gain important en matière de rapidité et de simplicité des preuves.[225] Quoi qu'il en soit, elle nous semble parfaitement remplir le cahier des charges de la (succincte) description du processus de réflexion donnée par Tennant (2002, 2005) ou Cieśliński (2010), à ceci près bien sûr qu'elle ne permet pas de justifier des schémas de réflexion suffisamment forts. Insistons sur ce point : si le processus de réflexion consiste à prendre conscience des engagements déductifs que nous avons implicitement contractés en acceptant T *au sens où*, ayant accepté tout énoncé prouvable dans T, nous devrions accepter tout énoncé dont on peut montrer qu'il est dérivable dans T, alors ce processus de réflexion ne fournit, nous semble-t-il, aucune raison d'accepter l'ensemble des énoncés d'implication du type

$$\mathrm{Rfn}(T) := Thm_T(\ulcorner \varphi \urcorner) \rightarrow \varphi$$

et encore moins du type

$$\mathrm{RFN}(T) := \forall x Thm_T(\ulcorner \varphi(\dot{x}) \urcorner) \rightarrow \forall x \varphi(x)$$

C'est ce qu'illustre le modèle procédural de l'acceptation : en termes de force logique, il y a bien un hiatus entre accepter tout énoncé en présence d'une preuve (dans T) de cet énoncé et même simplement en présence d'une preuve que cet énoncé est prouvable-

224. *Cf.* Beklemishev (2005, p. 210). En logique (modale) de la prouvabilité —où le carré modal $\Box_T$ est « interprété » comme un prédicat de prouvabilité (voyez Boolos (1993) pour plus de détails)— ce principe de réflexion énoncé comme *règle* est parfois noté :

$$\frac{\Box_T \varphi}{\varphi}$$

225. Voir Parikh (1971, 1973) pour des résultats de *speed-up* impliquant ce type de règle, ou bien Beklemishev (2005) ainsi que les indications bibliographiques qui s'y trouvent.

dans-T, et accepter l'ensemble infini des axiomes d'un *schéma* de réflexion.

Mais peut-être n'avons-nous pas rendu justice au processus de réflexion évoqué par Tennant. Peut-être qu'ayant accepté T, nous sommes non seulement engagés à accepter tous les théorèmes de T, non seulement engagés à accepter tous les énoncés φ tels qu'on peut prouver $Thm_T(\ulcorner\varphi\urcorner)$, mais que nous sommes encore engagés à accepter certains énoncés de la forme $Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$. Après tout, cette assertion énonce (peut-être) une relation d'implication entre le fait que φ est prouvable dans T et le fait que (nous acceptons (?)) φ. Elle est (peut-être) ainsi une manière d'énoncer la fiabilité de T par rapport à φ, ou bien encore, si on essaye de suivre les traces de Tennant et Cieśliński, une façon d'« exprimer notre confiance » en T concernant φ. L'assertion $Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$ constitue donc (peut-être) une manière d'articuler les raisons que nous avons d'accepter φ lorsque $T \vdash \varphi$, autrement dit une manière d'articuler ce à quoi nous nous sommes engagés lorsque nous avons accepté T.

Néanmoins, là encore, si Thm_T, le prédicat de prouvabilité dans T, satisfait un certain nombre d'hypothèses raisonnables, on a des résultats techniques qui limitent grandement la portée de cette autre manière de comprendre le processus de réflexion. En effet, le théorème de Löb montre que si on accepte les énoncés du type $Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$ uniquement pour les énoncés φ qui sont des théorèmes de T, on n'obtiendra là encore qu'une extension conservative de T. Plus précisément :

Théorème. (Löb, 1955) Soit T une théorie exprimée dans $\mathcal{L}_T$ et contenant assez d'arithmétique pour pouvoir exprimer sa propre syntaxe. Soit Thm_T un prédicat de prouvabilité pour T exprimé dans $\mathcal{L}_T$ et satisfaisant les trois conditions suivantes [226] :

$D1$. Si $T \vdash \varphi$, alors $T \vdash Thm_T(\ulcorner\varphi\urcorner)$

$D2$. $T \vdash Thm_T(\ulcorner\varphi \rightarrow \psi\urcorner) \rightarrow (Thm_T(\ulcorner\varphi\urcorner) \rightarrow Thm_T(\ulcorner\psi\urcorner))$

$D3$. $T \vdash Thm_T(\ulcorner\varphi\urcorner) \rightarrow Thm_T(\ulcorner Thm_T(\ulcorner\varphi\urcorner)\urcorner)$

alors pour tout énoncé $\varphi \in \mathcal{L}_T$, on a :

226. Dites conditions de dérivation de Löb-Hilbert-Bernays. (voyez le chapitre précédent). Le théorème de Löb est un résultat classique et bien connu. Pour une démonstration voir par exemple Smoryński (1977) ou Beklemishev (2005). Notez que si ce résultat est aujourd'hui parfaitement standard, il a néanmoins été considéré comme bien surprenant lorsqu'il est apparu (ainsi Boolos (1993, p. 54) le qualifie de « proprement stupéfiant pour au moins cinq raisons », voyez également Smoryński (1991) pour une discussion du contexte historique et intellectuel qui a accompagné la parution du résultat de Löb, ainsi qu'une discussion des subtiles distinctions entre les conditions de Hilbert-Bernays *stricto sensu* et celles de Löb).

$$\text{Si } T \vdash Thm_T(\ulcorner\varphi\urcorner) \to \varphi \text{ alors } T \vdash \varphi$$

Comme la direction inverse est évidente on a donc en fait, sous les mêmes hypothèses concernant Thm_T :

$$T \vdash Thm_T(\ulcorner\varphi\urcorner) \to \varphi \text{ ssi } T \vdash \varphi$$

Ce qui doit être clair à présent et qui est techniquement incontestable, c'est qu' « accepter »

(a) tous les φ tels que $T \vdash \varphi$

(b) tous les φ tels que $T \vdash Thm_T(\ulcorner\varphi\urcorner)$

(c) tous les φ tels que $T \vdash Thm_T(\ulcorner\varphi\urcorner) \to \varphi$

(d) tous les $Thm_T(\ulcorner\varphi\urcorner) \to \varphi$ pour φ tel que $T \vdash Thm_T(\ulcorner\varphi\urcorner)$ ou $T \vdash \varphi$

revient rigoureusement au même, en tout cas *du point de vue de la stricte force logique de ce qui est accepté*. Tous ces ensembles d'énoncés sont des extensions conservatives de PA et les accepter revient donc à accepter tous les théorèmes de PA... et rien d'autre. *A contrario* accepter l'ensemble infini des énoncés de la forme

$$Thm_T(\ulcorner\varphi\urcorner) \to \varphi$$

pour tout énoncé φ[227] de $\mathcal{L}_{PA}$, c'est adopter une théorie beaucoup plus forte.

Nous savions déjà que l'extension de T par l'ensemble infini de nouveaux axiomes

$$\text{Rfn}(T) : Thm_T(\ulcorner\varphi\urcorner) \to \varphi$$

(ou *a fortiori* par RFN(T)) produisait une extension stricte. Le théorème de Löb nous donne une mesure beaucoup plus fine et précise du phénomène : pour tout énoncé φ qui est un théorème de T (*i.e.* tel que $T \vdash \varphi$ *et* $T \vdash Thm_T(\ulcorner\varphi\urcorner)$), postuler $Thm_T(\ulcorner\varphi\urcorner) \to \varphi$ n'a, comme on pouvait s'y attendre, rigoureusement aucun effet en termes de renforcement de notre système de preuve[228]. À l'inverse, postuler $Thm_T(\ulcorner\varphi\urcorner) \to \varphi$ pour

227. y compris ceux sur lesquels PA ne « nous dit rien », au sens où elle ne les démontre ni ne les réfute.
228. Si $T \vdash \varphi$ alors trivialement $T \vdash Thm_T(\ulcorner\varphi\urcorner) \to \varphi$.

n'importe quel énoncé φ qui n' est *pas* un théorème de T produit une extension stricte de T [229]. Dans une telle situation, on a bien sûr $T \nvdash \varphi$, $T \nvdash Thm_T(\ulcorner\varphi\urcorner)$ (par K.) mais également $T \nvdash \neg Thm_T(\ulcorner\varphi\urcorner)$ [230].

Autrement dit, d'après le théorème de Löb, la non-conservativité de Rfn(T) sur T est entièrement et exactement due aux énoncés de la forme $Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$ pour lesquels T ne me donne aucune information sur l'antécédent $Thm_T(\ulcorner\varphi\urcorner)$ (T ne prouve ni $Thm_T(\ulcorner\varphi\urcorner)$ ni $\neg Thm_T(\ulcorner\varphi\urcorner)$), ni ne prouve le conséquent φ (tout au plus φ peut être réfutable : $T \vdash \neg\varphi$, ou bien être tout simplement indécidable dans T : $T \nvdash \varphi$ et $T \nvdash \neg\varphi$). Sous la seule hypothèse que j'ai accepté T, quelle raison puis-je avoir d'accepter également ce type d'énoncé ? La réponse ne semble pas évidente. Peut-être y a-t-il une manière d' « accepter » T d'après laquelle mon « acceptation » m'engage à une forme de « contrefactuels » :

Si T prouvait φ alors φ

y compris pour les énoncés que, justement, T ne prouve pas. Dans ce cas, si j'accepte la théorie de la relativité alors je suis implicitement engagé à « accepter » un énoncé tel que

(a) Si la théorie de la relativité prouve que la lune est faite de fromage vert alors la lune est faite de fromage vert.

De même, si j'accepte PA alors je suis implicitement engagé à accepter un énoncé tel que

(b) $Thm_{PA}(\ulcorner 1+1=55\urcorner) \rightarrow 1+1=55$

Nous ne voulons certainement pas dire ici qu'une telle notion d'« acceptation » est dénuée de toute plausibilité ou qu'elle est manifestement absurde [231]. Mais nous voulons

229. Pour reprendre la jolie formule attribuée à Parikh :

> PA ne pourrait pas être plus modeste concernant sa propre véracité (Boolos, 1993, cité p. 55).

230. En général, si T est cohérente et si Thm_T est défini de manière suffisamment naturelle, alors T ne prouve $\neg Thm_T(\ulcorner\varphi\urcorner)$ pour *aucun* énoncé φ. C'est une conséquence assez directe du second théorème d'incomplétude de Gödel. Sous les hypothèses du théorème de Löb, on peut donner une autre démonstration élégante de ce résultat : soit T une théorie cohérente, et soit Thm_T un prédicat de prouvabilité satisfaisant $D1 - D3$, alors soit φ un énoncé de $\mathcal{L}_T$. Si $T \vdash \neg Thm_T(\ulcorner\varphi\urcorner)$ alors trivialement $T \vdash Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$. Mais d'après le théorème de Löb, il s'en suit qu'alors $T \vdash \varphi$, d'où (par D1.) $T \vdash Thm_T(\ulcorner\varphi\urcorner)$, ce qui contredit la cohérence de T.

231. Bien au contraire, si par exemple j'accepte T au sens où je la tiens pour vraie, alors d'après les arguments de Ketland et Shapiro, je *suis* engagé envers ce type de « contrefactuels ».

simplement souligner que les raisons que nous pouvons avoir d' « accepter » une implication telle que (a) ou (b) nous semblent intuitivement plutôt résider dans le fait que nous avons des raisons de croire ou d'espérer (ou peut-être d'« accepter ») que l'antécédent de ces deux implications est faux [232]. Ce que montre le théorème de Löb, c'est que

1. les énoncés de ce type, *i.e.* ceux dont l'antécédent $Thm_T(\ulcorner\varphi\urcorner)$ est faux quoique T elle-même soit incapable de l'établir, sont précisément ceux qui sont pertinents pour la non-conservativité des schémas de réflexion.
2. les raisons que nous pouvons avoir pour notre conviction que $Thm_T(\ulcorner\varphi\urcorner)$ est faux, ne sont, semble-t-il, pas à chercher dans T elle-même puisque T est incapable de prouver $\neg Thm_T(\ulcorner\varphi\urcorner)$ (ni $Thm_T(\ulcorner\varphi\urcorner)$).

Ces raisons sont-elles à chercher dans notre « acceptation » de T ? Peut-être, mais d'un autre côté remarquons que l'on pourrait aussi défendre l'idée qu'une attitude raisonnablement prudente consisterait à accepter T et tous ses théorèmes, tout en restant résolument neutre au sujet des énoncés à propos desquels T ne nous dit rien, c'est-à-dire ceux qu'elle ne prouve pas et pour lesquels elle est incapable de nous dire s'ils sont ou non des théorèmes.

Pour poursuivre notre exploration des avatars de l'acceptation, voici en tout cas encore une forme d'acceptation qui *ne* semble *pas* valider ce type d'extension par un schéma d'axiomes appuyée sur un « processus de réflexion ». Supposons qu'au court de mon enquête scientifique, j'ai acquis la conviction qu'une sous-partie stricte de mon entreprise théorique est épistémiquement sûre tandis que le reste me semble beaucoup plus douteux. Soit T_0 la théorie formalisant la partie sûre, celle que j'accepte sans état d'âme ni hésitation. Supposons en outre que je souhaite néanmoins étendre T_0 en une théorie plus large T_1 (avec $T_0 \subsetneq T_1$) qui formalise la part douteuse de mon discours scientifique. Les raisons pour lesquelles je veux étendre T_0 à T_1 sont purement pragmatiques : je ne crois pas véritablement ce que semble dire $T_1 \setminus T_0$, et je n'entends certainement pas

232. Si nous « acceptons » (a) ou (b) c'est sans doute, nous semble-t-il, avant tout parce que nous croyons, espérons ou « acceptons » que
— la théorie de la relativité ne prouve pas que la lune est faite de fromage vert
— $\neg Thm_{PA}(\ulcorner 1+1=55\urcorner)$
Nous en voulons pour preuve le fait que si nous devions renoncer à cette conviction, c'est-à-dire si nous avions tout à coup des raisons de croire que la théorie de la relativité prouve que la lune est faite de fromage vert, ou que PA prouve que 1+1=55, nous en viendrions aussitôt à jeter aux orties ces théories. Mais quelles raisons pouvons-nous avoir de croire, « accepter » ou espérer que ces antécédents sont faux ? Elles ne sont visiblement pas à chercher dans le contenu de T elle-même. Peut-être est-ce parce que nous croyons que tous les théorèmes de T sont vrais ou corrects... ?

endosser les concepts et les objets douteux qui y apparaissent. Mais les outils fournis par T_1 facilitent grandement la manipulation des concepts et des objets de T_0 et simplifient amplement les calculs et démonstrations à leur sujet. Par bonheur, j'ai pu établir (en n'employant que des notions déjà accessibles dans T_0) un résultat de conservativité de T_1 sur T_0. Je suis donc assuré que le passage par T_1 ne m'induira pas en erreur concernant T_0, *i.e.* la part sûre de mon entreprise de connaissance. Fort de ce résultat j'accepte donc T_1 parmi mes méthodes (licites) de preuves, mais de manière purement instrumentale. Ici, le résultat de conservativité est capital pour justifier mon acceptation de T_1. Pour bien marquer ce dont il s'agit je dirai que j'*accepte*$_i$ T_1, avec un indice i pour instrumentalement [233].

J'ai donc *accepté*$_i$ T_1 au motif que T_1 est une extension conservative de T_0. Cette *acceptation*$_i$ n'est pas dénuée d'une certaine force : ainsi on peut dire que j'*accepte*$_i$ tous les théorèmes de T_1 [234]. On pourrait même stipuler que j'*accepte*$_i$ non seulement les théorèmes de T_1 mais également tous les énoncés dont je peux prouver (dans T_1) qu'ils sont des théorèmes de T_1 (*i.e.* tous les énoncés tels que $T_1 \vdash Thm_{T_1}(\ulcorner\varphi\urcorner)$ [235] Toutefois, cette *acceptation*$_i$ m'engage-t-elle implicitement à accepter (ou *accepter*$_i$) « à la réflexion » un schéma tel que $\mathrm{Rfn}(T_1)$? Le problème, c'est qu'il se peut très bien que, alors même que T_1 est une extension conservative de T_0, $T_1 \cup \mathrm{Rfn}(T_1)$ ne le soit plus [236] ! Dès lors, en ne faisant (d'après Tennant) qu' « expliciter mes engagements »

233. Bien évidemment, cet exemple est directement inspiré des conceptions hilbertiennes concernant les mathématiques finies/idéales. Pour fixer les idées, on peut prendre pour T_0 l'ensemble des mathématiques finitistes (disons $\mathcal{PRA}$, l'arithmétique primitive récursive), et pour T_1 une sous-partie de l'ensemble des mathématiques idéales pour laquelle on aurait pu fournir une preuve de conservativité sur T_0 (disons WKL_0). Voyez le chapitre précédent pour une exposition plus détaillée des conceptions de Hilbert et leurs liens avec la conservativité.

234. Du moins tous ceux exprimés dans $\mathcal{L}_{T_0}$. Mais on peut même imaginer que j'*accepte*$_i$ également tous les théorèmes de T_1 exprimés dans $\mathcal{L}_{T_1}$ si tant est qu'on puisse les considérer comme doués de signification. En tout cas, je les *accepte*$_i$ au moins à titre instrumental par exemple lorsqu'ils apparaissent comme une étape dans une preuve d'un énoncé de $\mathcal{L}_{T_0}$ donnée dans T_1.

235. Du moins si un tel mouvement résulte toujours en une extension conservative de T_0, ce qui sera le cas si Thm_{T_1} vérifie le critère $K1$ de Kreisel (voir 4.3.3.2).

236. Pour prendre un exemple trivial montrant que ce type de situation est possible, le rappel suivant devrait suffire :
Prenez
— $T_0 = PA$,
— $T_1 = PA_d = PA \cup \{Vr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in \mathcal{L}_{T_0}\}$.
Alors T_1 est une extension (déductivement) conservative de T_0. Mais $T_1 \cup \mathrm{Rfn}(T_1)$ ne l'est certainement pas.
Notez que,

1. bien que $\mathcal{L}_{T_1}$ contienne un prédicat de vérité pour $\mathcal{L}_{T_0}$, le schéma $\mathrm{Rfn}(T_1)$ s'énonce sans recourir à un prédicat de vérité *pour* $\mathcal{L}_{T_1}$. On demeure donc dans un cadre typé et évitons les problèmes

tacites contractés lorsque j'ai *accepté*$_i$ T_1, je scie en quelque sorte la branche sur laquelle je suis assis : en tout cas, je supprime la raison même (c'est-à-dire la conservativié sur T_0) sur laquelle reposait précisément mon *acceptation*$_i$ de T_1. Ainsi, l'application à T_1, théorie que j'ai *acceptée*$_i$, du processus de réflexion ou du raisonnement réflexif qui, pour reprendre les dires de Tennant lui-même, consiste simplement en

> une réflexion sur nos engagements axiomatiques et déductifs actuels, et une tentative d'articuler la conséquence systématique de ces engagements [...] (Tennant, 2005, p. 92)

de sorte qu'

> [a]ucune justification supplémentaire n'est nécessaire pour le nouvel engagement pris en exprimant nos précédents engagements [et que] dès l'instant qu'on apprécie à sa juste valeur le processus de réflexion [...] on possède déjà une explication des raisons pour lesquelles quelqu'un qui accepte [une théorie] devrait également accepter toutes les instances du principe de réflexion [pour cette théorie]. (Tennant, 2005, p. 92)

conduit tout bonnement ici à une contradiction méthodologique flagrante [237] !

Bien entendu, il y a certainement des « façons » d'« accepter » T_1 qui cautionnent son extension par $\mathrm{Rfn}(T_1)$: par exemple si j'accepte T_1 au sens où je la tiens pour vraie [238], si je la prends « au sérieux » ou « pour argent comptant » (*at face value* comme disent les anglo-saxons)...*etc.* Toutefois, l'*acceptation*$_i$ n'est visiblement pas un « mode » d'acceptation de cette sorte. Tennant pourrait peut-être objecter que l'*acceptation*$_i$ n'est pas vraiment une acceptation au sens où lui-même l'entend, que je ne me suis pas « véritablement engagé » envers T_1 ou bien encore que je ne lui fais pas « vraiment » confiance. Peut-être, peut-être pas : on pourrait répondre à cela que *tant qu'elle est conservative* on a autant confiance en T_1 qu'en T_0, et que si la notion d'*acceptation*$_i$ n'est certainement pas celle que Tennant veut invoquer dans sa stratégie, c'est à lui de nous en dire plus

de paradoxes du type menteur.

2. si le codage de la syntaxe de $\mathcal{L}_{T1}$ et de $\mathcal{L}_{T_0}$ n'est pas trop pathologique, $T_1 \cup \mathrm{Rfn}(T_1)$ prouve $Con(T_1)$ et donc *a fortiori* $Con(T_0)$.
3. il existe sans aucun doute des exemples mathématiquement plus intéressants de ce type de phénomène (on peut penser à l'extension de PA par ACA_0...).

237. Pour une argumentation différente mais aboutissant à une conclusion similaire selon laquelle un hilbertien *ne* devrait *pas* accepter les schémas de réflexion, voyez Dean (2015).
238. auquel cas, je pourrais peut-être même fournir une dérivation de ce schéma de réflexion en m'appuyant, entre autres, sur certaines propriétés du concept de vérité.

sur ce qu'il faut comprendre précisément lorsqu'il utilise ce terme. On ne peut donc que regretter que Tennant ne soit pas plus prolixe au sujet de la notion d'acceptation qu'il a en vue et sur laquelle il entend s'appuyer pour justifier l'adoption de schémas de réflexion. Si l'on veut proposer une justification de ces schémas à partir d'une analyse des « engagements » que nous prenons ou sommes censés prendre lorsque nous « *acceptons* » une théorie, s'en remettre à une compréhension intuitive, pré-théorique ou pré-formelle du (ou d'un) concept d'acceptation nous semble en réalité largement insuffisant.

Mais plutôt que de déplorer ce qui paraît manquer (lourdement) à la discussion d'une justification des principes de réflexion à partir de la notion d'acceptation telle qu'elle a été présentée par Tennant (2002, 2010, 2005), concentrons-nous plutôt sur les intéressants développements auxquels elle a donné lieu.

4.3.3.3 La critique de la prémisse (3)/(F)

Nous avons déjà signalé que Ketland (2010) trouvait à redire à l'argument de Tennant-Cieśliński (exposé page 363). Il questionne notamment passage de (1) à (2). Toutefois, pour Ketland (2010), c'est la prémisse (3)/(F), qu'il appelle *Principe d'expression*, qui est la plus problématique. Ketland (2010) baptise d'ailleurs *stratégie expressioniste* l'argumentation de Tennant revue et corrigée par Cieśliński. On peut en effet se demander en quel sens exactement le principe de réflexion Rfn(PA) peut être dit « exprimer » le contenu de (D).

Ketland observe tout d'abord qu'

> [...] il s'agit là d'une thèse non-standard. Habituellement, un schéma de réflexion tel que Rfn(S) est dit exprimer la fiabilité[239] de S : que tout ce que S prouve est *vrai*. Et être *vrai* n'est pas la même chose qu'être *accepté*. (Ketland, 2010, p. 430)

Vers la fin de son article, il souligne en outre l'importance critique d'expliquer plus précisément ce qu'il faut entendre par « exprimer » ici puisque c'est sur cette notion que repose crucialement la stratégie de Tennant et Cieśliński. En guise d'analyse de la relation « X exprime une partie du contenu de Y », laissée inexpliquée dans Cieśliński (2010), Ketland formule la suggestion suivante (Ketland, 2010, p. 434), :

$$X \text{ exprime une partie du contenu de } Y \text{ ssi } Y \text{ implique } X.$$

239. En anglais : *soundness*, qu'on pourrait également traduire par correction.

Selon cette suggestion, il y a bien un sens à soutenir que Rfn(PA), le principe de réflexion local, « exprime » une partie du contenu du principe de réflexion globale $Ref(PA)$ puisque, nous l'avons vu, sous réserve qu'on prenne une théorie axiomatisant le prédicat de vérité suffisante pour obtenir les **T**-équivalences , $Ref(PA)$ implique Rfn(PA) [240]. De manière similaire, pour soutenir l'idée que RFN(PA), le principe de réflexion uniforme, « exprime » une partie du contenu de $Ref(PA)$, il faudra s'appuyer sur des hypothèses aléthiques auxiliaires un peu plus fortes, suffisamment fortes pour dériver les **T**-équivalences uniformes [241].

À l'inverse, comme le fait remarquer Ketland (2010), si on accepte cette analyse de la relation « X exprime le contenu de Y », la prémisse (3) devient

(3') : (D) implique Rfn(PA)

Mais, poursuit Ketland (2010, p. 434), cette dernière assertion semble clairement fausse. Et la stratégie expressioniste est donc vouée à l'échec.

Aux yeux d'un logicien, il ne fait aucun doute que la relation « X implique Y » semblera plus familière, plus claire, et plus appréhendable que la locution « X exprime le contenu de Y ». La notion d' « expression partielle d'un contenu » peut apparaître vague, insuffisamment précise et définie, ou à tout le moins en attente d'une clarification, en particulier lorsque —comme c'est le cas ici— X et Y sont des ensembles d'énoncés exprimés dans des langages différents. À l'inverse, la relation d'implication, du moins s'il faut entendre par là la relation d'implication déductive, est le pain quotidien du logicien. De ce point de vue, le contenu de (3') peut paraître plus facilement évaluable que celui, quelque peu mystérieux, de (3), et le remplacement de (3) par (3') un progrès dans l'analyse de l'argument. Toutefois, si nous sommes tentés à première vue de partager le diagnostic négatif de Ketland quant à la plausibilité de (3'), il nous semble qu'en réalité on n'a fait que déplacer le problème. Il est en effet capital de prendre conscience ici que la valeur d'une assertion du type « Y implique X » va dépendre de manière décisive des hypothèses auxiliaires et du cadre théorique d'arrière-plan dans le-

240. On a donc :

Rfn(PA) « exprime une partie du contenu de » $Ref(PA)$ puisque $Ref(PA) \vdash_{PA_d}$ Rfn(PA).

241. Voyez la section 4.3.1 sur les diverses formulations de la réflexion et les relations de déducibilité reliant celles-ci. On constatera ici que la notion d' « ... est une expression partielle du contenu de ... », si tant est qu'on l'examine selon les lignes proposées par Ketland, peut être à géométrie variable en fonction de la théorie d'arrière plan, ou des hypothèses auxiliaires qu'on s'autorise à employer pour l'analyser.

quel s'inscrit cette assertion.[242] Ceci sera d'autant plus important lorsque X et Y sont des ensembles d'énoncés formulés dans des langages différents : si les vocabulaires de X et Y sont distincts, X et Y ne pourront être dans un rapport d'implication (déductive) qu'au moyen de ce qu'on pourrait appeler —sur le modèle, par exemple, de l'analyse du rôle des termes théoriques en relation au contenu empirique d'une théorie scientifique— des « lois ponts » reliant le contenu de X et celui de Y.[243] Ainsi, pour statuer précisément sur la question de savoir si (D) implique ou n'implique pas Rfn(PA), c'est-à-dire sur la question de savoir si Rfn(PA) « exprime » une partie du contenu de (D),[244] il est indispensable de se demander quelles *ressources auxiliaires* on peut mobiliser pour tenter de dériver Rfn(PA) à partir de (D). Après tout, peut-être y a-t-il une théorie τ formalisant notre acceptation de PA[245] qui soit suffisamment forte pour sanctionner (3'), c'est-à-dire qui soit suffisamment forte pour relier le vocabulaire de (D) et celui de Rfn(PA) et mettre au jour des lois ponts entre ces deux ensembles d'énoncés, de sorte que Rfn(PA) devienne effectivement déductible de (D) modulo τ.[246]

242. Ce point est déjà illustré par le cas des extension aléthiques et du principe de réflexion global dans la mesure où lorsqu'on prend comme axiomatisation du prédicat de vérité la théorie décitationnelle limitée aux **T**-équivalences classiques, c'est-à-dire PA_d^{+ind}, on aura que $Ref(PA)$ implique Rfn(PA) puisque

$$PA_d^{+ind} \cup Ref(PA) \vdash Thm_T(\ulcorner\varphi\urcorner) \rightarrow \varphi \text{ pour tout énoncé } \varphi \in \mathcal{L}_{PA},$$

i.e. $Ref(PA)$ associé à PA_d^{+ind} permet de dériver toutes les instances de Rfn(PA), tandis que $Ref(PA)$ n'implique pas RFN(PA) puisque

$$PA_d^{+ind} \cup Ref(PA) \nvdash \forall x Thm_T(\ulcorner\varphi(\dot{x})\urcorner) \rightarrow \forall x \varphi(x) \text{ pour toute formule à une variable libre } \varphi(x) \in \mathcal{L}_{PA}.$$

Nous l'avons déjà dit : pour dériver toutes les instances de RFN(PA) à partir de $Ref(PA)$, il faut une théorie aléthique d'arrière-plan aussi forte que $PA_{d.u.}$, *i.e.* la décitation *uniforme*.

En d'autres termes, les ensembles d'énoncés de $\mathcal{L}_{PA}$ dont on peut dire qu'ils « expriment » (au sens proposé par Ketland) une partie du contenu du principe de réflexion global $Ref(PA)$, vont *varier selon la théorie aléthique d'arrière-plan considérée*.

243. Un cas limite étant celui où le vocabulaire d'un des membres de la relation « X implique Y » est réductible à celui du second, par exemple au moyen de définitions explicites et d'interprétations relatives d'une théorie dans une autre. Ces cas sont bien connus des logiciens, des mathématiciens et des philosophes des sciences. Ils ont notamment leur importance en philosophie de la physique. Pour une entrée générale sur ces questions nous renvoyons à Batterman (2012). Pour un exemple plus proche du sujet discuté ici et bien connu des mathématiciens, on peut également penser à la réduction de la notion d'entier à celle d'ensemble par définition et lois ponts (quelque problématique et discutée qu'elle soit). Dire que PA exprime une partie du contenu de ZF, c'est-à-dire que ZF implique PA, n'aura de sens qu'une fois qu'on aura expliqué comment les énoncés de l'arithmétique se traduisent (modulo définitions ou stipulations) par des énoncés du langage de la théorie des ensembles.

244. Toujours selon la suggestion de Ketland (2010).

245. Et dans laquelle, bien sûr, aucune notion substantielle ou explicative de vérité n'interviendrait.

246. Tout comme la dérivation des principes de réflexion uniforme ou locale à partir de $Ref(PA)$ dépend d'une théorie aléthique d'arrière-plan.

Bien sûr, aux yeux de Ketland (2010), l'existence d'une telle théorie τ de l'acceptation est extrêmement douteuse et intuitivement peu plausible. Et c'est pourquoi, il peut déclarer que (3') semble « clairement fausse ». Mais c'est au fond tout le problème : même une fois la relation « Rfn(PA) exprime une partie du contenu de (D) » analysée et réduite au moyen de la notion d'implication, la question demeure de savoir quelle théorie d'arrière plan, et plus spécifiquement en l'occurrence quelle théorie formalisant notre acceptation de PA, on peut mobiliser dans l'analyse d'une éventuelle relation d'implication entre (D) et Rfn(PA). Nous sommes d'avis qu'une exigence légitime minimale qu'on peut adresser à un partisan de l'argument de Tennant-Cieśliński est de donner bien plus d'indications sur ce à quoi pourrait ressembler une telle théorie τ, autrement dit de clarifier la notion d'acceptation permettant de justifier (3') qu'il a en tête.[247] C'est précisément ce défi qu'ont tenté de relever Cieśliński (2017) et d'autres partisans du déflationnisme tels que Galinon (2010) et Horsten et Leigh (2017).

4.3.3.4 Intermezzo : une dérivation de Rfn(PA) ?

En première analyse, si on tente d'imaginer ce que pourraient être précisément les possibles rapports d'implication entre (D) et Rfn(PA), il faut sans doute commencer par une formalisation plus poussée de (D). Informellement, (D) s'énonçait par « si φ est un théorème de PA, alors je suis prêt à accepter φ ». En s'appuyant sur des ingrédients déjà disponibles (codage dans PA de sa propre syntaxe, définition d'un prédicat de prouvabilité), on peut proposer la formalisation suivante :

$$\text{(D')} : \forall x(Thm_{PA}(x) \rightarrow Acc(x))$$

où « $Acc(x)$ » est un *nouveau* symbole de prédicat unaire au moyen duquel nous avons enrichi notre langage de base $\mathcal{L}_{PA}$ pour obtenir un langage $\mathcal{L}_\tau \supseteq \mathcal{L}_{PA} \cup \{Acc(x)\}$ contenant les ressources lexicales suffisantes pour développer une théorie de l'acceptation, théorie qu'on pourra note τ. L'interprétation attendue de « $Acc(x)$ » sera l'ensemble des (codes des) énoncés de $\mathcal{L}_{PA}$ que j'accepte —ou peut-être...que je « devrais accepter en principe », ou bien que « je suis disposé à accepter ou à asserter », ou bien encore que « je me suis implicitement engagé à accepter », *etc.* , *etc.* ... selon le concept non-aléthique

247. Ou bien encore, les deux démarches n'étant pas exclusives l'une de l'autre, d'expliquer pourquoi (3') n'est pas une bonne explication de (3) et d'en donner une analyse alternative. Il nous semble toutefois incontestable qu'*il faut* plus d'explications.

et pragmatique d' "accepter" ou d' "asserter" un énoncé que l'on souhaitera mettre en avant pour parvenir effectivement à (D') [248] puis à une justification de Rfn(PA), concept dont τ est censée être une théorie.

Une première remarque consiste à noter que le détour par un langage et une théorie enrichis au moyen d'un *nouveau* symbole de prédicat $Acc(x)$ semble indispensable. En effet, si (D') est censé (avec le concours de τ) *impliquer* (ou du moins justifier) Rfn(PA), alors $\tau \cup (D')$ doit aboutir à une extension non-conservative de PA. Il s'en suit que (D') (ou plus exactement $\tau \cup (D')$) ne saurait être une extension définitionnelle de PA et que $\tau \cup (D')$ doit renfermer des concepts non explicitement définissables [249] dans PA et son langage. [250] Par conséquent, si τ est une théorie de l'acceptation $Acc(x)$ (et rien d'autre) on ne saurait avoir de formule $\Phi(x)$ de $\mathcal{L}_{PA}$ telle que

$$PA \vdash \forall x(Acc(x) \leftrightarrow \Phi(x))$$

248. Nous reprenons ici les formulations de Cieśliński (2010, Section 3.2)

> [...] je supposerai simplement que je peux effectivement arriver à (D) sans employer aucun concept de vérité (mais uniquement le concept pragmatique d' "accepter" ou d' "asserter" un énoncé). [...]

Voyez précédemment page 363.

249. ni même implicitement définissables (au sens habituel où l'extension est fixée de manière unique dans tout modèle) puisque nous sommes en premier ordre et que le théorème de Beth s'applique.

250. Ceci, simplement parce que $\tau \cup (D')$ n'est pas conservative sur PA et qu'une extension définitionnelle est toujours conservative : sur les liens entre extensions définitionnelles et extensions conservatives d'une théorie voir Hodges (1993, Section 2.6, et en particulier exercices 6, 7 et 8 p. 66).

soit vérifiée.[251] Une autre conséquence de ceci valant la peine d'être relevée est que la notion d'acceptation qui se cache derrière le prédicat $Acc(x)$, puisqu'elle ne peut pas être définissable dans PA, devra aussi nécessairement être distincte de la prouvabilité dans PA.

Ces remarques étant formulées, supposons donc que nous raisonnions à partir de

$$\text{(D')} : \forall x(Thm_{PA}(x) \rightarrow Acc(x))$$

accompagné éventuellement d'autres énoncés de $\mathcal{L}_\tau$ composant une théorie τ axiomatisant « $Acc(x)$ ». Comment obtenir $\text{Rfn}(PA)$? Doit-on espérer dériver $\text{Rfn}(PA)$ à partir de $(D') \cup \tau$? Faut-il fournir une preuve en bonne et due forme du type $(D') \cup \tau \vdash \text{Rfn}(PA)$? Par sa forme logique (D') est très proche du principe (sémantique) de réflexion global. (D') possède même rigoureusement la même forme syntaxique que $Ref(PA)$: $\forall x(Thm_{PA}(x) \rightarrow Vr(x))$, à ceci près qu'on aurait simplement remplacé « Vr » par « Acc ». Pour *dériver* $\text{Rfn}(PA)$ à partir de (D'), on pourrait donc être tenté de s'inspirer du raisonnement employé dans la démonstration sémantique de $\text{Rfn}(PA)$ à partir de

251. Nous disons que τ doit être une théorie de l'acceptation $Acc(x)$ et *rien d'autre*, car bien sûr il faut bien poser certaines limites pour être sûrs que c'est bien la notion $Acc(x)$ qui fait le travail dans la dérivation de $\text{Rfn}(PA)$. *A contrario*, voici un exemple pathologique illustrant ce que nous voulons dire : supposons que, cherchant à montrer que (D) implique $\text{Rfn}(PA)$ modulo τ, on développe une théorie τ axiomatisant $Acc(x)$ mais contenant également un certain nombre de ressources auxiliaires, sur lesquelles on n'impose pas de limites. Supposons plus précisément que $\mathcal{L}_\tau$ contienne également un prédicat « $Vr(x)$ » et que $PA^{+ind}_{Tar} \subseteq \tau$. Alors trivialement, on aura bien évidemment $\tau \cup (D) \vdash \text{Rfn}(PA)$ quel que soit le contenu de (D), pour la bonne et simple raison qu'on a déjà $\tau \vdash \text{Rfn}(PA)$. Dans un tel cas pathologique, on pourra prendre une définition fantaisiste pour $Acc(x)$ dans $\mathcal{L}_{PA}$. Par exemple, $Acc(x) :=_{def} \exists y(x = y)$, ou bien $Acc(x) :=_{def} Thm_{PA}(x)$. On aura bien alors

$$(D) : \forall x(Thm_{PA}(x) \rightarrow Acc(x))$$
$$\text{et } \tau \cup (D) \vdash \text{Rfn}(PA)$$

On voit à nouveau les difficultés que soulève l'évaluation précise d'une thèse telle que

$$\text{Rfn}(PA) \text{ « exprime » une partie du contenu de (D).}$$

Même si on l'analyse selon les lignes esquissées par Ketland et qu'on l'interprète par

$$\text{(D) implique Rfn}(PA),$$

de nombreux problèmes et ambiguïtés demeurent. D'un côté, en l'absence de toutes hypothèses annexes, comme une théorie d'arrière-plan τ formalisant les concepts contenus dans (D) et des lois ponts reliant ces concepts avec ceux de $\text{Rfn}(PA)$, l'existence d'un rapport d'implication entre (D) et $\text{Rfn}(PA)$ est virtuellement impossible (ne serait-ce que parce que (D) et $\text{Rfn}(PA)$ ne sont pas formulés dans le même langage). D'un autre côté, si aucune limite n'est placée sur la théorie d'arrière-plan τ que l'on peut mobiliser pour tenter de dériver $\text{Rfn}(PA)$ à partir de (D), alors on peut se trouver dans des cas limites comme ci-dessus où $\text{Rfn}(PA)$ est trivialement dérivable sans pour autant qu'il y ait le moindre sens à affirmer que $\text{Rfn}(PA)$ exprime le contenu de (D).

$Ref(PA)$ [252] .

Par simple instantiation on obtient facilement :

$$(D_1) \quad Thm_{PA}(\ulcorner\varphi\urcorner) \rightarrow Acc(\ulcorner\varphi\urcorner) \text{ pour tout } \varphi \in \mathcal{L}_{PA}$$

Pour un énoncé $\varphi \in \mathcal{L}_{PA}$, l'instance de (D_1) correspondante nous dit que si φ est un théorème alors φ est accepté. Étant donné la forme syntaxique de (D_1) et celle de $\text{Rfn}(PA) : Thm_{PA}(\ulcorner\varphi\urcorner) \rightarrow \varphi$, l'axiome (ou plutôt le schéma d'axiomes) suivant :

$$(*) \quad Acc(\ulcorner\varphi\urcorner) \rightarrow \varphi \text{ pour } \varphi \in \mathcal{L}_{PA}$$

semble avoir exactement la forme logique adéquate pour être le chaînon manquant permettant de relier (D_1) à $\text{Rfn}(PA)$. Ce schéma d'axiome jouerait alors un rôle strictement parallèle à celui des **T**-équivalences dans la dérivation sémantique du principe de réflexion uniforme. [253] *Grosso modo*, $(*)$ relie l'acceptation [254] d'un énoncé φ au contenu de cet énoncé lui-même et affirme que si φ est accepté alors φ « doit être le cas » ; pour le dire moins maladroitement, $(*)$ nous dit que si φ est accepté alors φ est vrai. [255]

Néanmoins, contrairement au cas de la dérivation sémantique de $\text{Rfn}(PA)$ dans une extension aléthique, nous ne pouvons pas ici nous appuyer *a priori* sur des axiomes décitationnels permettant d'annuler « $Acc(x)$ » et de remplacer $Acc(\ulcorner\varphi\urcorner)$ par φ. La

252. Tennant (2002) lui-même insistait sur le fait qu'il serait souhaitable d'imiter la structure formelle des arguments sémantiques lorsqu'on veut en trouver des équivalents déflatés acceptables pour les déflationnistes.

253. Voici mises en parallèle les deux dérivations :

Dérivation sémantique		Dérivation par l'acceptation	
1. $\forall x(Thm_{PA}(x) \rightarrow Vr(x))$	$(Ref(PA))$	1. $\forall x(Thm_{PA}(x) \rightarrow Acc(x))$	(D)
2. $Thm_{PA}(\ulcorner\varphi\urcorner) \rightarrow Vr(\ulcorner\varphi\urcorner)$	inst. de $\forall$	2. $Thm_{PA}(\ulcorner\varphi\urcorner) \rightarrow Acc(\ulcorner\varphi\urcorner)$	inst. de $\forall$
2.′ $Vr(\ulcorner\varphi\urcorner) \rightarrow \varphi$	**T**-équivalence	2.′ $Acc(\ulcorner\varphi\urcorner) \rightarrow \varphi$	$(*)$(« décitation » pour Acc)
3. $Thm_{PA}(\ulcorner\varphi\urcorner) \rightarrow \varphi$	2 et 2' + transitivité de $\rightarrow$	3. $Thm_{PA}(\ulcorner\varphi\urcorner) \rightarrow \varphi$	2 et 2' + transitivité de $\rightarrow$

Pour parodier la phraséologie déflationniste, on pourrait dire qu'il s'agit ici d'« annuler la montée vers l'acceptation » au moyen d'une (demie) **Acc**-équivalence $(*)$, à l'instar de la façon dont la décitation « $Vr(\ulcorner\varphi\urcorner) \rightarrow \varphi$ » est censée annuler la montée sémantique...

254. Ou l' « acceptation implicite », ou l' « engagement à asserter », ou *etc.* , selon le concept pragmatique (et non aléthique) que le prédicat « $Acc(x)$ » est censé désigner.

255. Et ce n'est pas qu'une façon de parler : modulo les **T**-équivalences , $Acc(\ulcorner\varphi\urcorner) \rightarrow \varphi$ et $Acc(\ulcorner\varphi\urcorner) \rightarrow Vr(\ulcorner\varphi\urcorner)$ sont en effet interdérivables. Il y a donc bien un sens à dire que $Acc(\ulcorner\varphi\urcorner) \rightarrow \varphi$ « dit » que si φ est accepté alors φ est vrai.

question de la plausibilité d'un principe tel que $(*)$ se pose donc de manière impérieuse. Un tel principe est-il acceptable ? Est-il justifié, et si oui par quelles raisons ? À première vue, *en toute généralité* pour une théorie T et un langage $\mathcal{L}_T$ quelconques un principe tel que

$$Acc(\ulcorner \varphi \urcorner) \rightarrow \varphi \text{ pour } \varphi \in \mathcal{L}_T$$

semble des plus discutables, pour ne pas dire franchement douteux, pour ne pas dire carrément inacceptable. Le fait que φ soit accepté par un agent cognitif —ou par une communauté de scientifiques, ou par un agent idéal parfaitement rationnel, *etc.* — ne devrait pas, selon nous, avoir *en général* pour conséquence que φ (est vrai).

Bien entendu, les considérations développées ci-dessus s'appuient seulement sur le fait que $(*)$ est un schéma d'axiomes *suffisant* pour dériver $\mathrm{Rfn}(PA)$ à partir de (D_1) (ou de (D')). Rien de ce que nous avons dit ne montre que $(*)$ est *indispensable* pour dériver $\mathrm{Rfn}(PA)$ à partir de (D'). Peut-être que d'autres principes peuvent jouer ce rôle. Peut-être même que ces autres principes pourraient donner une justification de $\mathrm{Rfn}(PA)$ à partir de (D') sans montrer également au passage que pour tout énoncé φ,

$$Acc(\ulcorner \varphi \urcorner) \rightarrow \varphi$$

c'est-à-dire avoir $(*)$ pour corollaire, ce qui serait une conséquence intuitivement peu plausible et problématique à nos yeux. Peut-être enfin que, contrairement à ce que suggère l'analyse de la « stratégie expressiviste » développée par Ketland (2010) et qui aboutit à exiger que

$$\text{(3') : (D) ou (D') } \textit{implique } \mathrm{Rfn}(PA),$$

y a-t-il une façon de justifier un principe de réflexion à partir d'une théorie de l'acceptation qui ne prenne pas la forme d'une déduction du type $(D') \cup \tau \vdash \mathrm{Rfn}(PA)$. Nous allons à présent examiner les tentatives formulées en ce sens au nom du déflationnisme.

4.3.3.5 Un première tentative : Galinon (2014, 2010) et la responsabilité épistémique

Face à ces difficultés, pour tenter de combler le fossé entre l'acceptation d'un énoncé et son contenu (ou entre l'acceptation d'une théorie et sa correction), pour tenter, en somme de rendre plus crédibles les implications $(*)$, ou plus généralement une inférence du type :

$$\begin{array}{c}(D)\\ \vdots\\ \hline \mathrm{Rfn}(PA)\end{array}$$

ou plus généralement encore, la possibilité d'une justification d'un schéma tel que $\mathrm{Rfn}(PA)$ à partir de la simple acceptation de PA, on pourrait proposer de renforcer les conditions portant sur la notion d'acceptation désignée par $Acc(x)$. On pourrait ainsi être tenté de parler d'acceptation justifiée, ou d'acceptation correcte ou légitime, ou bien encore d'acceptabilité, et autres choses du même type. Ceci permettra en effet d'évoquer certaines contraintes supplémentaires s'appliquant à l'acceptation d'une théorie, qui pourront à leur tour être invoquées dans le cadre d'une justification des principes de réflexion. Bien entendu, on devra prendre garde de ne pas réintroduire subrepticement ce faisant une notion substantielle ou explicative de vérité. [256]

Les analyses développées par Henri Galinon (2010, chapitre 6, p. 215-244) et reprises dans Galinon (2014) nous semblent pouvoir être classées parmi ce type d'argumentations. Elles sont certes un peu différentes du raisonnement que nous avons esquissé ci-dessus à la suite de Tennant (2002) revu par Cieśliński (2010) (et par Ketland (2010)). Cependant, elles nous semblent très proches en esprit.

Très proche en esprit dans les intentions, tout d'abord : à l'instar de Tennant (2002), Galinon se propose de montrer que

> [...] si un sujet accepte une théorie donnée A, alors il *doit* accepter que A est cohérente, et cela pour des raisons relevant de sa compréhension de sa propre activité théorique, [257] indépendamment du contenu de la théorie qu'il accepte et de toute réflexion sur la notion de vérité. (Galinon, 2010, p. 216, italiques de l'auteur)

Là encore, il s'agit de s'appuyer sur une analyse des engagements pris par quelqu'un qui accepte une théorie de façon à justifier une extension de cette théorie sans passer par la notion de vérité. Pour reprendre une autre formulation donnée dans Galinon (2014), il s'agit de montrer que

> l'acceptation de la cohérence par un agent rationnel qui accepte une théorie donnée est justifiée *par défaut*, sur la base d'un certain nombre de principes

256. Par exemple, en stipulant par construction qu'une théorie acceptable est vraie...
257. Cette « compréhension de sa propre activité théorique » est peut-être à rapprocher du « processus de réflexion » cher à Tennant.

> qui relèvent purement de la rationalité en première personne. (Galinon, 2014, p. 321)

Très proche en esprit dans les moyens et dans la forme, ensuite : pour justifier le renforcement d'une théorie de base T qu'un agent aurait acceptée, Galinon se propose de s'inspirer de la dérivation de la cohérence de T obtenue à partir du principe (sémantique) global de réflexion pour T. Sur le modèle de ce principe global de réflexion $Ref(T)$, qu'il rebaptise Principe de réflexion aléthique, Henri Galinon définit ce qu'il appelle un

> **Principe de réflexion épistémique** : T est acceptable [258]

Si on identifie T à l'ensemble de ses théorèmes, alors ce principe de réflexion épistémique devrait se formaliser par l'énoncé suivant :

$$(D^*) : \forall x(Thm_T(x) \rightarrow Acc^*(x)) \text{ (Principe de réflexion épistémique)}$$

Dans sa forme, cet énoncé nous est maintenant familier, et on peut le rapprocher de l'énoncé qui est apparu lorsque nous avons tenté de formaliser (D) dans l'argument de Tennant-Cieśliński. Mais *ici, « $Acc^*(x)$ » est cette fois interprété par une notion d'acceptation* justifiée *ou selon les termes de* Galinon *(2014, 2010) par une notion d' «* acceptabilité *»*, ce qui est certainement une notion plus forte que la simple acceptation *de facto* d'une théorie par un sujet ou un agent cognitif. Cette nuance est donc fondamentale et nous allons y revenir.

Par parenthèse, remarquons toutefois préalablement qu'une autre distinction que l'on peut relever entre les analyses de Galinon (2014, 2010) et l'argument de Tennant-Cieśliński réside dans le fait que, du moins dans Galinon (2010, chapitre 6, p. 215-244) Galinon se cantonne à l'objectif de justifier l'acceptation de la cohérence de T plutôt

258. Voyez Galinon (2010, p. 235). Dans Galinon (2014, p. 327), la définition du principe de réflexion épistémique proposée est un peu différente puisqu'elle prend cette fois la forme suivante :

> **Principe de réflexion épistémique** : Je suis justifié à accepter T,

ce qui se formaliserait, toujours en identifiant T à l'ensemble de ses théorèmes, de la manière suivante :

$$(D^{**}) : \forall x(Thm_T(x) \rightarrow J(x))$$

où $J(x)$ est interprété par un prédicat signifiant « je suis justifié à accepter x ». Mais en fait les deux formulations sont plus ou moins synonymes aux yeux de l'auteur, puisque pour lui la notion d'acceptabilité d'une théorie est elle-même dérivée du fait que je suis justifié à l'accepter :

> En effet, que signifie l'affirmation que je suis justifié à accepter la proposition ou la théorie T, *ou comme je dirai de façon synonyme*, que T est acceptable (par moi, maintenant) ? [...] (Galinon, 2014, p. 327, nous soulignons)

que d'un principe de réflexion uniforme ou local.[259] Cependant, si la construction qu'il donne concerne la justification de l'énoncé de la cohérence, Henri Galinon semble lui-même considérer que sa démarche doit pouvoir se généraliser aux principes de réflexion, comme l'atteste l'extrait suivant :

> Les logiciens, en particulier Solomon Feferman (1991, 1962), ou John Myhill (1960), qui se sont intéressés aux « principes de réflexion », du type « Si p est prouvable alors p », ont reconnu depuis longtemps que ces principes s'imposent rationnellement à quelqu'un qui est engagé dans la pratique de la démonstration par certains moyens de preuve, mais ils ont surtout cherché à étudier la force logique de ces principes – ce qui s'en déduit – et se sont peu intéressés à leur justification. *On peut voir l'appel au principe de responsabilité comme un premier pas dans cette direction.*
>
> (Galinon, 2014, p. 331, nous soulignons)

Ceci le rapproche encore un peu plus de la démarche mise en avant par Tennant et Cieśliński.

Mais revenons à la justification de la cohérence. Nous avons dit que la notion d'acceptabilité, ou d'acceptation justifiée était sans nul doute une notion plus forte que l'acceptation *tout court*. Bien entendu, il appartient à Galinon de préciser quelle est cette notion d'acceptabilité qu'il a en vue. Dans Galinon (2014, 2010) il fournit quelques éléments d'analyse. Tout en se référant à la littérature philosophique portant sur la notion d'acceptation[260], il souligne en premier lieu que l'acceptation doit être clairement distinguée de la croyance. Contrairement à cette dernière, l'acceptation, au sens où l'entend Henri Galinon, est « volontaire[261] ». C'est même un « acte réfléchi[261] ». Elle est « spécifiquement le produit d'un examen critique[261] » et obtenue à l'issue d' « un processus de délibération rationnelle[261] ». À ce titre, l'acceptation est certainement la notion la mieux à même de décrire l'attitude épistémique que doit entretenir l'homme ou la femme de science vis-à-vis du produit final de son activité théorique, dans la mesure où celui-ci doit pouvoir être rationnellement contrôlé. Ainsi, accepter au sens développé ici « procède d'une décision réflexivement informée et guidée uniquement par des buts qui sont ordi-

259. Et nous avons vu que les principes de réflexion produisent en général des extensions strictement plus fortes que celles obtenues au moyen de la seule cohérence. La cohérence est donc une barre a priori plus facile à atteindre.
260. Spécifiquement : Cohen (1989), Engel (2000) et Van Fraassen (1980).
261. Ces qualifications sont tirées de Galinon (2010, p. 224)

nairement reconnus pour être ceux de l'activité scientifique [262] ». Galinon met ensuite en avant plusieurs propriétés remarquables de cette notion d'acceptation qui ont une importance pour notre discussion. Tout d'abord, l'acceptation —contrairement (peut-être ?) à la croyance— n'est pas nécessairement soumise à un impératif de vérité. Rien ne nous contraint par exemple à considérer que l'acceptation vise la vérité [263]. D'autre part, dans la mesure où l'acceptation procède d'un acte volontaire, elle engage notre responsabilité. Elle est en particulier soumise à certaines normes de rationalité semblables à celles qui existent en théorie de la décision. En reprenant une expression fameuse de Clifford (1877), Henri Galinon considère ainsi qu'il existe une « éthique de l'acceptation », un sorte de code épistémique auquel un individu *rationnel* devrait se conformer lorsqu'il *décide* d'accepter. Accepter (rationnellement) une théorie nous rend donc comptables de certaines justifications. Avec cette notion forte d'acceptation, soumise à des contraintes de rationalité, Henri Galinon peut introduire une notion d'acceptabilité (ou d'acceptation justifiée) à l'œuvre dans son principe de réflexion épistémique :

> T est acceptable [...] signifie que mon acceptation de T satisfait à une certaine norme, que j'affirme qu'il m'est permis, au regard d'un certain code tacite d' « éthique épistémique », d'accepter T. (Galinon, 2010, p. 235) [264]

La question de savoir ce que contient exactement un tel « code de conduite épistémique rationnelle » est évidemment bien difficile. Aux dires de Galinon (2014, p. 327), elle est aussi difficile que celle de la nature de la justification elle-même [265]. Malgré cela, d'après Galinon (2014, 2010), il y a deux propriétés saillantes de la notion d'acceptabilité sur lesquelles on peut s'appuyer.

La première est que

> l'acceptabilité n'a rien à voir avec la vérité, ce sont des normes différentes. (Galinon, 2010, p. 235)

Ce point est évidemment fondamental ici, c'est-à-dire dans le cadre d'une défense du déflationnisme, puisque, on l'aura compris, il s'agit de donner une justification de la cohérence de T (et peut-être également d'un principe de réflexion pour T) qui ne fasse

262. Tiré de Galinon (2014, p. 321).
263. Un exemple paradigmatique particulièrement célèbre d'une relation d'acceptation qui ne consiste pas à tenir pour vrai et auquel Galinon (2010) se réfère, est celui développé par Van Fraassen (1980).
264. Cette formulation est églaement reprise *verbatim* dans (Galinon, 2014, p. 327).
265. Galinon (2014) renvoie ensuite le lecteur à Alston (1988) pour une entrée dans la vaste littérature portant sur la conception déontique de la justification.

aucun appel, même subreptice, même implicite à une notion de vérité [266].

La seconde est que parmi les règles qui gouvernent *a priori* la notion d'acceptabilité, c'est-à-dire parmi les conditions sous lesquelles nous sommes justifiés à accepter ce que nous acceptons, se trouve minimalement l'exigence que l'ensemble des énoncés que nous acceptons soit cohérent [267]. Et cette prescription peut, là encore, être obtenue sans lien aucun avec la notion de vérité, mais uniquement à partir de contraintes pragmatiques présidant à notre activité de théorisation :

> Le problème n'est pas seulement qu'une théorie incohérente doive être fausse (car après tout, à nouveau, cette idée n'a qu'une application limitée pour un instrumentaliste). Le problème est qu'une théorie incohérente est inutile. [...]
>
> Par conséquent il est hautement plausible que, de même qu'une analyse conceptuelle de la notion de vérité révèle que l'ensemble des énoncés vrais est cohérent, de même une analyse conceptuelle de la notion de justification ou

266. D'ailleurs, cette distinction radicale entre acceptabilité et vérité pourrait tout à fait être contestée. Certes, il semble à peu près indiscutable que l'acceptabilité, *i.e.* les conditions sous lesquelles nous pouvons nous considérer comme justifiés à accepter une théorie ou un énoncé, et la vérité sont probablement extensionnellement distinctes. Il existe sans doute des énoncés/des théories vrai(e)s que nous ne sommes pas justifiés à accepter, *i.e.* qui ne sont pas acceptables, tout comme il y a certainement des théories acceptables aujourd'hui qui pourront se révéler fausses. Mais peut-on pour autant affirmer que les deux notions n'ont rien à voir l'une avec l'autre ? Ne pourrait-on pas dire par exemple que la justification vise la vérité ? Que la norme ultime de justification est la vérité et que, dès lors une théorie ultimement acceptable devra être vraie ? Sommes-nous absolument certains que la justification et/ou l'acceptabilité peuvent être définies ou même simplement analysées d'une manière qui soit totalement indépendante de toute notion de vérité ? Nous n'en sommes pas persuadés. Mais c'est évidemment une question sérieuse dont l'étude détaillée dépasse de beaucoup les limites du présent travail. Nous n'en pousserons donc pas plus loin l'examen. Notre grief avec les arguments de Galinon est d'une autre nature comme on s'en apercevra ci-dessous (*cf. infra*).

267. Ce second point pourrait lui aussi être contesté. Sommes-nous si certains que parmi les injonctions de notre code de bonne conduite épistémique ne figurent pas des principes contradictoires ? Par exemple, on peut considérer

1. qu'il est justifié/rationnel d'accepter notre meilleure théorie scientifique actuelle.

De même, sans doute faut-il admettre

2. qu'il est justifié/rationnel d'accepter tout énoncé expérimentalement vérifié.

Mais alors que penser de Newton qui, *en son temps*, acceptait une théorie physique dont on sait *aujourd'hui* qu'elle était fausse et même contredite par des résultats expérimentaux *de nos jours* parfaitement établis ? Dynamiquement, nous avons simplement amendé notre meilleure théorie physique. Mais *au temps de Newton*, d'un point de vue si l'on peut dire synchronique, 1. et 2. pouvaient en fait mener à une contradition $p \wedge \neg p$...
Nous pouvons le voir aisément aujourd'hui, mais *en son temps*, Newton lui-même en était incapable. Faut-il conclure qu'il n'était pas rationnel, qu'il acceptait des propositions injustifiées ? La cohérence de l'ensemble des énoncés que nous sommes justifiés à accepter ne va donc peut-être pas aussi de soi que semble le penser Henri Galinon. Ceci dit, nous laisserons également cette seconde question de côté.

> d'acceptabilité doit révéler que si un agent est justifié à accepter un ensemble d'énoncés, ou si un ensemble d'énoncés est acceptable [...], alors cet ensemble d'énoncés est également cohérent. (Galinon, 2010, p. 236)

Dès lors, une fois munis de cette notion forte d'acceptabilité on peut, selon Galinon (2010, p. 237), obtenir

> une dérivation de la cohérence de T à partir du principe de réflexion épistémique qui est tout à fait analogue à la dérivation de la cohérence de T *à partir* du principe de réflexion aléthique. Au lieu de faire appel à des lois *a priori* de la vérité, néanmoins, cette dérivation ne fait appel qu'à une analyse élémentaire des buts qui gouvernent l'action d'accepter telle que nous l'avons présentée [...]

Cette dérivation prend la forme suivante (Galinon, 2010, p. 237) [268] :

1. T est acceptable (Principe de Réflexion épistémique)
2. Si T est acceptable, alors T est cohérente (réflexion sur les normes d'acceptabilité)
3. Donc T est cohérente (par 1, 2)

À ce stade, on pourrait se demander en quoi la dérivation ci-dessus est « tout à fait analogue » à la dérivation de la cohérence de T à partir du principe $(Ref(T))$. Quels liens ou quelles similarités y a-t-il entre l'analyse conceptuelle de la notion de justifiction ou d'acceptabilité qui sous-tend la prémisse 2. ci-dessus et la dérivation en bonne et due forme de l'énoncé de la cohérence $Con(T)$ (ou de l'énoncé G_T) dans une extension aléthique suffisamment forte ? La dérivation de G_T à partir de (Ref) fait par exemple un usage important des **T**-équivalences . En écho à nos interrogations précédentes, on pourrait donc se demander si l'analyse conceptuelle de l'acceptabilité devrait nous amener à adopter un principe tel que

$$(**) \quad Acc^*(\ulcorner\varphi\urcorner) \rightarrow \varphi \text{ pour } \varphi \in \mathcal{L}_T, \text{ où } Acc^*(x) \text{ signifie « } x \text{ est acceptable » } [269]$$

Cette question est toutefois secondaire et, de toute façon, les éléments avancés dans Galinon (2014, 2010) ne sont pas suffisamment formalisés pour qu'une réponse évidente

268. Repris *verbatim* dans Galinon (2014, p. 328).
269. Cette interrogation sera particulièrement pressante si, comme Galinon le suggère lui-même, la notion d'acceptabiltié doit pouvoir être employée pour justifier non pas seulement la cohérence de T mais carrément un principe de réflexion pour cette théorie.

se dégage. Nous la laisserons donc de côté. Ainsi, admettons qu'à partir du principe de réflexion épistémique pour T, *i.e.* de la prémisse 1. ci-dessus, on puisse obtenir une justification raisonnablement convaincante de la cohérence de T.

Pour autant nous ne sommes pas au bout de nos peines. Si nous voulons bien admettre qu'une théorie acceptable est cohérente, en quoi cela nous permet-il de montrer qu'un agent (fût-il rationnel) qui accepte une théorie de base T est justifié, ou bien même simplement engagé, à accepter que T est cohérente ? Et en quoi cela nous fournit-il une justification de la cohérence de T susceptible de remplacer la dérivation sémantique de $Con(T)$ et de G_T [270] ? Après tout le fait qu'un agent accepte une théorie n'a certainement pas pour conséquence que celle-ci est acceptable. La pièce manquante de l'édifice nous est donnée par Galinon sous la forme de ce qu'il appelle un principe de responsabilité :

> Le principe qui permet de faire le pont entre la petite dérivation [ci-dessus] et cette idée qu'un sujet acceptant une théorie donnée doit accepter que cette théorie est cohérente est le suivant [...] :
>
> **Principe de Responsabilité** : Si un agent rationnel S accepte un ensemble d'énoncés X, S doit accepter « X est acceptable ».
>
> *Si* ce principe est correct, en effet, nous avons l'explication cherchée :
>
> 1. S accepte T (notre hypothèse de départ)
> 2. Donc S doit accepter « T est acceptable » (par 1 et Responsabilité)
> 3. Or S doit juger que si X est acceptable, alors X est cohérent (réflexion sur les normes d'acceptabilité/justification)
> 4. Donc S doit accepter « T est cohérent » (2 et 3)
>
> (Galinon, 2014, p. 328-329)

Il est important de comprendre que la justification ou l'explication proposée par Galinon (2014, 2010) est d'une nature fondamentalement différente de celle mise en avant par Ketland (1999) et Shapiro (1998b). Henri Galinon en a parfaitement conscience et le revendique d'ailleurs ouvertement puisqu'il déclare :

> La justification de *l'acceptation de la cohérence* [271] par le principe de responsabilité est d'une nature fondamentalement différente. Ce n'est pas une

270. Ou *a fortiori* d'un principe de réflexion, local, uniforme ou autre ?
271. Nous soulignons.

> *preuve*[272] de la cohérence : un indice que j'accepte T *n'*est *pas*[272] , sans hypothèses substantielles supplémentaires, un *indice*[272] , même indirect, du fait que T est acceptable et de la cohérence de T. Cette justification du caractère rationnel de l'acceptation de la cohérence d'une théorie que nous acceptons (aussi longtemps que nous l'acceptons) est donc une forme de justification *par défaut*[272], qui vaut en l'absence d'indice de la vérité de la cohérence. C'est aussi une justification *défaisable*[272] [...] Elle n'en est pas moins rationnelle. (Galinon, 2014, p. 331-332)

Ainsi, nous ne sommes pas en présence d'une preuve de la cohérence de T, donnée sous la forme d'une dérivation logique à partir d'un certain nombre d'axiomes. La conclusion, ou le point d'arrivée, du raisonnement ci-dessus n'est pas l'énoncé de la cohérence lui-même. C'est au contraire un énoncé exprimant la nécessité ou le devoir pour un agent rationnel S ayant accepté une théorie T d'accepter également que cette théorie est cohérente.[273] C'est donc un énoncé portant sur les croyances de S (ou plus précisément sur les énoncés acceptés par S) vis-à-vis de T et non pas (ou pas directement) sur le contenu ou sur les propriétés logiques (comme par exemple la cohérence) de T elle-même. Il énonce plutôt certaines contraintes ou certains impératifs de rationalité s'exerçant sur les « acceptations » d'un agent S. Ce qu'établit ou tente d'établir le raisonnement ci-dessus n'est donc pas la cohérence de T *stricto sensu* mais plutôt la nécessité pour un agent rationnel ayant accepté T d'accepter également la cohérence de cette théorie.

En outre, l'argumentation développée dépend crucialement du Principe de Responsabilité en première personne. La question est donc de savoir si ce principe est correct. Henri Galinon fournit une variété d'arguments à l'appui de ce principe, les plus développés se trouvant dans Galinon (2010). Il évoque tout d'abord les travaux de Burge (1996) et de Wright (2004a,b), d'après lesquels (ne serait-ce que pour éviter une régression à l'infini) il existe des propositions dont l'acceptation est « justifiée » ou plutôt « autorisée » (*entitled*)[274] par défaut, c'est-à-dire en l'absence de toute justification obte-

272. Italiques de l'auteur.

273. Pour une version plus formalisée de cette dérivation, qui met peut-être mieux en lumière la nature de sa conclusion, voyez l'annexe **??**.

274. Ces auteurs anglo-saxons distinguent parmi les garanties épistémiques (*epistemic warrants*), d'une part les justifications proprement dites (*justifications*) (obtenues après un argument, une preuve, une observation, *etc.*) et les permissions ou droits ouverts (*entitlements*) qui nous autorisent à postuler sans justification supplémentaire un certain nombre de propositions fondamentales sans lesquelles aucune entreprise de connaissance ne pourrait démarrer ni être fondée. Sur l'introduction de ces deux types de garanties épistémiques, voyez par exemple Burge (1996, p. 93).

nue à partir de principes plus fondamentaux. Ces propositions sont ce que Wright appelle les « pierres de touches » de toute entreprise cognitive, des hypothèses sans lesquelles nous ne pourrions regarder aucune de nos méthodes de justification comme correctes. Parmi ces propositions, se rangent sans doute certaines lois fondamentales de la logique, ou des énoncés du type :

Je ne suis pas le jouet d'un malin génie.

Selon Galinon (2014, 2010), le principe de responsabilité en première personne, et donc de manière dérivée la cohérence de ce que nous acceptons elle-même, est à rapprocher de ces « pierres de touches » mises en avant par Wright et Burge.

Pour appuyer et développer un peu plus son propos, Galinon (2010) offre une intéressante comparaison avec le « paradoxe » de Moore (1942), rendu particulièrement célèbre par Wittgenstein (2004, II.x p. 190). Ce type de paradoxe concerne l'absurdité apparente qu'il y a pour un individu à affirmer (ou croire, ou juger, ou accepter) *simultanément* et *en première personne* les deux conjoints d'un énoncé de la forme suivante :

Il pleut & je ne crois pas qu'il pleut. [275]

Il est assez communément admis qu'un individu rationnel ne devrait pas accepter simultanément que p et qu'il ne croit pas que p. Et il est aussi assez communément admis que ce paradoxe ne repose pas, ou pas directement, sur une contradiction logique. Les deux conjoints ci-dessus sont parfaitement compatibles d'un point de vue strictement logique et la conjonction qu'ils forment n'est pas une antilogie. Bien plus, je peux tout à fait croire ou affirmer qu'il pleut à un moment donné ; je peux tout à fait croire ou affirmer que je ne crois pas qu'il pleut à un moment donné. Les énoncés « il pleut » et « je ne crois pas qu'il pleut » peuvent tout à fait être simultanément vrais. Mais si *moi-même*, je crois ou affirme ces deux énoncés en même temps, alors je dis ou fais quelque chose d'absurde, ou du moins d'irrationnel. Il s'agit bien d'un paradoxe *à la première personne* qui met en jeu le rapport que nous pouvons avoir à nos propres croyances ou aux énoncés que nous acceptons.

Moore lui-même avait ainsi déjà observé que le paradoxe disparaît dès lors que l'on change le temps ou la personne dans la conjugaison du verbe apparaissant dans le second

275. Ce qui est la version dite omissive du paradoxe (p et $\neg$ (je crois que p)). La version commissive étant : « Il pleut et je crois qu'il ne pleut pas » (*i.e.* p et je crois que ($\neg p$)). Cette distinction et ces appellations des différentes versions du paradoxe sont devenues courantes depuis l'étude fondatrice de ce problème par Hintikka (1962). Pour une synthèse récente des travaux sur le paradoxe de Moore et ses liens avec la rationalité, voyez Green et Williams (2007).

conjoint. Rien d'absurde à affirmer qu'il pleuvait mais que je ne croyais pas qu'il pleuvait. Rien de problématique à considérer qu'il pleut mais que tu ne crois pas qu'il pleut. Le paradoxe n'apparaît que lorsqu'un même agent forme ou énonce une première croyance et, au même instant, forme ou énonce une seconde croyance (inverse) à propos de cette première croyance qui lui est propre. Dans une telle situation une forme de dissonance cognitive apparaît, qui contredit le caractère de transparence de nos propres pensées à nous-mêmes. Si je crois que p, sans doute dois-je croire et même savoir [276] que je crois que p. À tout le moins, je dois ne pas croire simultanément que je ne crois pas que p. Au-delà de la stricte cohérence logique apparaissent donc des principes de rationalité qui imposent une forme de concordance entre ce qu'on pourrait appeler mes croyances de premier ordre (*i.e.* des croyances à propos du monde) et ce qu'on pourrait appeler mes croyances de second ordre (*i.e.* des croyances à propos de mes propres croyances).

De la même manière, selon Galinon, il serait absurde pour un individu d'accepter simultanément une théorie T et une proposition niant que T soit acceptable. Ceci serait absurde non pas au sens où cela serait logiquement incohérent [277], mais au sens où, à l'instar des énoncés de Moore, cela placerait aussitôt l'individu dans une situation paradoxale du point de vue de sa propre rationalité. Pour illustrer cette idée, Henri Galinon examine l'exemple suivant :

> [imaginons] le cas d'un agent rationnel qui accepterait :
>
> (*) La terre tourne autour du soleil, mais je ne suis pas justifié à accepter que la terre tourne autour du soleil. [278]

Bien que les conditions de vérité de cet énoncé (*) ne soient pas problématiques, un agent rationnel qui accepterait cet énoncé se placerait aussitôt dans une situation « moo-

276. On pourrait objecter ici qu'il existe peut-être des formes de croyances inconscientes. Je pourrais alors croire qu'il pleut mais ignorer cependant cette croyance et même croire (à tort) que je ne crois pas qu'il pleut. Nous n'ouvrirons pas ce dossier ici. Disons simplement que, d'une part, si on parle non plus de croyance mais d'acceptation, considérée comme une décision rationnelle obéissant à délibération réfléchie, alors la possibilité d'une acceptation rationnelle inconsciente semble vraiment ténue, voire tout bonnement impossible par définition. D'autre part, si on s'interroge sur la dimension normative de l'acceptation d'une théorie (c'est-à-dire sur les engagements implicites que nous contractons lors de cette acceptation) dans le cadre de l'activité scientifique, sans doute devrait-on laisser de côté les cas « pathologiques » d'acceptation ou de croyance inconscientes, si tant est que de telles choses existent et aient un sens.

277. Puisque bien sûr la cohérence (ni sans doute son acceptabilité) de T n'est pas une conséquence logique de T elle-même.

278. Exemple tiré de (Galinon, 2010, p. 238)

réeenne ». En effet, croire (ou accepter) pour un agent rationnel qu'il n'est pas justifié à accepter que la terre tourne autour du soleil [279] devrait immédiatement s'accompagner d'un abandon ou d'une modification de la croyance (ou de l'acceptation) que la terre tourne autour du soleil [280]. Manquer de le faire, ce serait faillir à nos impératifs de rationalité, ce serait échouer à suivre notre « éthique de l'acceptation ». Pour Henri Galinon, le point commun de ces décisions d'acceptation conduisant à des paradoxes à la Moore est

> qu'elles violent un *principe de responsabilité constitutif de notre rationalité* : les croyances constituées dans l'examen réflexif de nos propres croyances et les croyances examinées doivent d'une certaine façon être *coordonnées* pour former un ensemble rationnel. [...]
>
> la *rationalité* d'un sujet commande que la nature de l'articulation de ses jugements à propos de ses propres jugements, avec ces derniers jugements eux-mêmes, incorpore essentiellement le fait que les uns comme les autres sont *ses* pensées. On n'a pas la même relation épistémologique, en termes de droits comme en termes de devoirs, avec le contenu de ses propres pensées et avec le contenu des pensées d'autrui, quand bien même ces contenus seraient identiques d'un point de vue sémantique. (Galinon, 2010, p. 238-239, italiques de l'auteur)

Ce qui se traduit par le Principe de responsabilité en première personne qui prend la forme d'une implication dont l'antécédent (l'hypothèse de départ) est bien qu'un agent S accepte une théorie T, et dont le conséquent *n*'est *pas*, bien entendu, que T est justifiée/acceptable, mais plutôt un énoncé affirmant l'impératif rationnel pour S d'*accepter* (par défaut ?) que T est acceptable/justifiée.

Ainsi, Galinon (2014, 2010) dresse un tableau plausible à l'appui du principe de responsabilité en première personne. Malgré tout, nous pensons que ce principe de responsabilité est incorrect. Bien plus, nous pensons être dans une certaine mesure en capacité de proposer une réfutation pure et simple de ce principe. Plus précisément, nous montrons que le principe de responsabilité, dès lors qu'on l'adjoint à un autre principe qui lui nous semble assez incontestable et que nous baptisons principe de prudence, dé-

279. C'est la croyance ou acceptation de « second ordre ».
280. c'est-à-dire de la croyance de « premier ordre » sur laquelle porte la croyance de second ordre et avec laquelle elle doit être rationnellement coordonnée.

bouche sur des conséquences inacceptables [281]. L'argument que nous proposons nécessite un niveau de formalisation un peu plus poussé. Reprenons pas à pas le raisonnement de Galinon (2014, 2010). Après avoir introduit un

Principe de Responsabilité : Si un agent rationnel S accepte un ensemble d'énoncés X, S doit accepter « X est acceptable ».

Galinon propose la dérivation suivante

> *Si* ce principe est correct, en effet, nous avons l'explication cherchée :
>
> 1. S accepte T (notre hypothèse de départ)
> 2. Donc S doit accepter « T est acceptable » (par 1 et Responsabilité)
> 3. Or S doit juger que si X est acceptable, alors X est cohérent (réflexion sur les normes d'acceptabilité/justification)
> 4. Donc S doit accepter « T est cohérent » (2 et 3)
>
> (Galinon, 2014, p. 328-329)

dont le résultat est bien qu'un agent ayant accepté une théorie T *doit* accepter que T est cohérente.

Comme l'auteur lui-même le reconnaît, l' « explication cherchée » repose crucialement sur le Principe de Responsabilité. Pour justifier ce principe, nous avons vu qu'Henri Galinon avance diverses considérations portant sur l'analyse de la rationalité et la responsabilité qu'induit pour nous la transparence de nos propres jugements. Nous n'y reviendrons pas ici [282]. Toutefois, nous pensons que le Principe de Responsabilité est incorrect. Pour le montrer, il est nécessaire de formaliser de manière un peu plus poussée les notions entrant en jeu dans le raisonnement de Galinon (2014, 2010). Nous introduisons donc les notations suivantes :

— Soit $A(x)$ un prédicat unaire signifiant « j'accepte x ».

— Soit $J(x)$ un prédicat signifiant « x est justifié (*i.e.* acceptable) » [283].

A et J peuvent s'appliquer à des énoncés ou à des théories, c'est-à-dire des ensembles éventuellement infinis d'énoncés. Dans ce qui suit nous nous limiterons au cas où A et J

281. Nous proposons donc une forme de *reductio* du principe de responsabilité.
282. Pour un exposé et une discussion plus détaillés, voyez page 392.
283. Les notions de théorie « acceptable » ou « justifiée » sont employées de manière à peu près synonyme par Henri Galinon (voyez la note 258). Nous notons cette double notion par $J(x)$ simplement pour la distinguer clairement de la notion d'acceptation *tout court* notée $A(x)$, et non pas pour privilégier la notion de justification par rapport à celle d'acceptabilité.

s'appliquent à des (noms ou des codes d') énoncés ou, si l'on veut, à des théories finies dont on aurait pris la conjonction des membres pour former un seul énoncé. Cela est sans importance pour notre argument et facilite sa formalisation. En particulier, la cohérence d'un énoncé (ou d'une théorie finie se réduisant à un énoncé) peut se formuler simplement en disant que cet énoncé n'est pas une contradiction (disons $\varphi \neq \bot$). Signalons également que nous identifierons les énoncés et leurs noms (ou leur code) pour faciliter la lecture, puisque ce point est également sans importance pour notre argument. De plus, les principes ci-dessous seront donnés sous la forme de schémas d'axiomes où φ apparaît comme une sorte de métavariable désignant un énoncé. Le langage de ces énoncés n'est pas précisé, il s'agira du langage de la théorie de base T [284] qu'un agent rationnel aurait acceptée. Par précaution, on pourra supposer que les prédicats A et J ne font pas partie de ce langage, ceci pour éviter les risques de paradoxes du type menteur. Mais en fait cela n'a pas d'importance non plus pour notre argument. Par ailleurs, Henri Galinon fait également usage dans son raisonnement de notions déontiques, telles que la notion de devoir épistémique, d'éthique de l'acceptation ou d'impératifs de rationalité. [285] Pour formaliser ces notions, nous emploierons la notation habituelle de logique déontique : le carré modal $\Box$ traduit l'obligation. Dans le contexte de notre discussion,

$\Box\varphi$ peut ainsi se lire comme « il est (rationnellement) impératif que φ ».

La nature exacte de cet impératif, le sens précis qu'Henri Galinon attache au verbe « devoir » apparaissant dans le Principe de Responsabilité, n'a pas d'importance pour notre argument. Simplement, lorsque nous aurons à faire à des énoncés du type « je *dois* (en tant qu'agent rationnel) faire ceci ou cela », nous les formaliserons par « $\Box$ (faire ceci ou cela) » [286].

Dès lors, ces précisions étant apportées, les principes employés **??** par Henri Galinon s'énoncent et se formalisent comme ceci :

Principe de responsabilité

Si un agent rationnel accepte un énoncé φ alors il doit accepter « φ est justifié » (*i.e.* $J(\varphi)$).

Formellement :

284. Théorie qui dans notre formalisation sera donc identifiée à un énoncé.
285. Voyez en particulier l'occurrence du verbe devoir dans la prémisse 2. de l' « explication cherchée ».
286. De manière (un peu) plus grammaticale : « $\Box$ (je fasse ceci ou cela) ».

$$Resp : A(\varphi) \rightarrow \Box A(J(\varphi))$$

Selon Galinon (2014, 2010) un examen de mon « éthique de l'acceptation », ou une analyse conceptuelle de la justification, nous donne le principe suivant :

Propriété de l'acceptabilité

Si φ est justifié (acceptable) alors φ n'est pas une contradiction

Formellement :

$$PAc : J(\varphi) \rightarrow (\varphi \neq \bot)$$

D'autre part, nous aurons à faire usage d'un autre principe sous-entendu dans la démonstration proposée par Galinon, (et plus ou moins explicitement énoncé dans la version plus détaillée donné dans Galinon (2010, p. 238)). Ce principe est nécessaire pour arriver à 4. à partir de 2. et 3. ; il concerne le fait que l'ensemble des énoncés que je dois accepter est clos déductivement :

Propriété de clôture de l'acceptation impérative

Si je dois accepter φ et si $\varphi \rightarrow \psi$ alors je dois accepter ψ

Formellement :

$$\Box\text{-clos} : [\Box A(\varphi) \wedge (\varphi \rightarrow \psi)] \rightarrow \Box A(\psi)$$

Avec ces outils formels en main, l' « explication cherchée » se formalise comme ceci (où pour plus de lisibilité, nous avons réécrit à chaque étape le principe employé dans le pas inférentiel, précédé d'un $*$, tout en conservant la numérotation d'origine) :

1.	$A(\varphi)$	*Fait*
$*$	$A(\varphi) \rightarrow \Box A(J(\varphi))$	*Resp*
2.	$\Box A(J(\varphi))$	par 1. et Resp
3.	$J(\varphi) \rightarrow (\varphi \neq \bot)$ [287]	*PAc*
$*$	$[\Box A(J(\varphi)) \wedge (J(\varphi) \rightarrow (\varphi \neq \bot))] \rightarrow \Box A(\varphi \neq \bot)$	$\Box$-clos
4.	$\Box A(\varphi \neq \bot)$	par 2., 3. et $\Box$-clos

C'est une déduction logiquement correcte. Comme attendu, sa conclusion n'est pas l'énoncé de la cohérence de φ. C'est au contraire une proposition affirmant l'obligation

287. Peut-être avons-nous simplifié ici. Ne faudrait-il pas plutôt formaliser 3. par :

$$\Box A(J(\varphi) \rightarrow (\varphi \neq \bot))$$

puisque Henri Galinon écrit : « *S doit* juger que ... » ? Cela n'a à vrai dire guère d'importance pour notre argument. Signalons simplement que si l'on voulait formaliser 3. comme ceci, il faudrait également adapter le principe $\Box$-clos pour obtenir une dérivation correcte formalisant l'argument de Galinon.

qui s'impose (pour un agent rationnel) d'accepter l'énoncé de la cohérence de φ. La dérivation de cette proposition sous l'hypothèse que j'ai accepté φ fournit, selon Galinon (2014, 2010) une forme de justification par défaut de l'acceptation de la cohérence de φ.

Mais nous avons un problème avec le Principe de Responsabilité en première personne qui relie ou mélange, en position d'antécédent, une hypothèse factuelle et, en position de conséquent, une affirmation déontique sur ce qui devrait être ou sur ce qui devrait s'imposer à moi (en tant que sujet rationnel). En guise de réfutation de ce principe, nous proposons l'argument suivant.

Tout d'abord, il existe un autre principe reliant la notion de justification/acceptabilité et les normes de rationalité, qui nous semble parfaitement défendable. Il nous semble même bien plus défendable que *Resp*, et ce point est de *la plus haute importance* pour notre argument. Nous le baptiserons :

Principe de prudence

Si φ n'est pas justifié (acceptable) alors il n'est pas vrai que je doive accepter que φ est justifé.

Formellement :

$$Pru : \neg J(\varphi) \rightarrow \neg \Box A(J(\varphi))$$

Remarquez que ce principe *ne* dit *pas* que

1. si φ n'est pas justifiée, alors je dois ne pas accepter que φ est justifiée ;

   ce qui s'énoncerait :

   $$\neg J(\varphi) \rightarrow \Box \neg A(J(\varphi))$$

   et qui est (peut-être) un principe plausible [288], mais en tout cas distinct du principe de prudence.

2. si φ n'est pas justifiée, je dois ne pas accepter φ ;

   ce qui s'énoncerait :

   $$\neg J(\varphi) \rightarrow \Box \neg A(\varphi)$$

   Après tout, peut-être y a-t-il des énoncés qui ne sont pas justifiés mais dont l'acceptation est néanmoins permise, *i.e.* des énoncés tels que $\neg J(\varphi) \wedge \neg \Box \neg A(\varphi)$),

288. Encore qu'il existe peut-être des énoncés non justifiés mais dont il est néanmoins permis d'accepter qu'ils sont justifiés.

(avec $\neg\Box\neg$ interprété par il n'est pas obligatoire de ne pas ..., *i.e.* il est permis de ...)

3. si φ n'est pas justifiée, je ne dois pas accepter φ ;

   ce qui s'énoncerait :

   $$\neg J(\varphi) \rightarrow \neg\Box A(\varphi)$$

   Après tout, peut-être y a-t-il des énoncés qui ne sont pas justifiés mais dont l'acceptation est néanmoins obligatoire, *i.e.* des énoncés φ tels que $\neg J(\varphi) \wedge \Box A(\varphi)$.

4. si φ n'est pas justifiée, alors je dois accepter $\neg\varphi$;

   ce qui s'énoncerait :

   $$\neg J(\varphi) \rightarrow \Box A(\neg\varphi)$$

   et qui est certainement faux, puisqu'il existe sans doute des énoncés qui ne sont pas justifiés sans que pour autant il soit rationnellement impératif d'en accepter la négation.

Pru dit seulement que si φ n'est pas justifié, alors il n'est pas rationnellement impératif que j'accepte (l'énoncé affirmant) que φ est justifié (c'est-à-dire en l'occurrence que j'accepte un énoncé faux).[289] Comme nous l'avons dit, ce principe de prudence nous semble donc tout à fait défendable (en tout cas, autant si ce n'est plus que le principe de responsabilité).

Le problème est que *Pru* et *Resp* sont en réalité incompatibles, car, ensemble, ils permettent la dérivation suivante :

289. *Pru* peut d'ailleurs se voir comme un cas particulier, cantonné au cas des énoncés portant sur la justification, d'un principe plus général :

$$PruGén : \neg\varphi \rightarrow \neg\Box A(\varphi)$$

lequel principe dit en substance que si φ est faux, il n'est pas rationnellement impératif que j'accepte φ. Autrement dit, je ne suis pas tenu d'accepter des énoncés faux. Notez également que ce principe équivaut (moyennant une loi de contraposition et d'élimination des doubles négations) au principe suivant :

$$PruGén' : \Box A(\varphi) \rightarrow \varphi$$

qui n'est pas sans rappeler un principe de réflexion et qui affirme que si je dois impérativement, si je suis rationnellement tenu d'accepter φ alors φ est vrai.

1. $\neg J(\varphi) \rightarrow \neg\Box A(J(\varphi))$	*Pru*
2. $A(\varphi) \rightarrow \Box A(J(\varphi))$	*Resp*
3. $\neg\Box A(J(\varphi)) \rightarrow \neg A(\varphi)$	2. contraposée
4. $\neg J(\varphi) \rightarrow \neg A(\varphi)$	1. et 3.
5. $A(\varphi) \rightarrow J(\varphi)$	4. contraposée à nouveau

Mais 5., qui affirme en substance que si j'accepte φ alors φ est justifié, est sans conteste absurde. Ce principe semble avaliser une forme grossière de *wishful thinking*. D'ailleurs, Galinon lui-même le reconnaît puisqu'il déclare :

> Pour comprendre ce qui est en jeu, il est important de noter que le principe suivant, avec son implication matérielle, est évidemment *faux* :
>
> J'accepte $X \rightarrow X$ est acceptable [290]
>
> Il peut être *vrai* qu'un agent rationnel accepte de fait la théorie A, sans que pour autant A satisfasse aux critères d'acceptabilité. C'est une situation banale dans laquelle l'agent s'est simplement trompé et une illustration parmi d'autres du fossé qui existe entre ce qui est et ce qui doit être. (Galinon, 2014, p. 329, italiques de l'auteur)

Nous ne saurions mieux dire. Il faut donc semble-t-il renoncer au Principe de Responsabilité en première personne, ou au Principe de Prudence, (ou bien encore, à quelques lois de la logique classique...) En ce qui nous concerne, dans la mesure où le principe de responsabilité nous semble (beaucoup) plus douteux que le principe de prudence, nous tâcherons de rester prudemment irresponsables. Et, dans la mesure où le principe central sur lequel elle s'appuie nous semble erroné, l'argumentation d'Henri Galinon nous paraît donc non probante. Y a-t-il d'autres façons d'enrichir et d'analyser la notion d'acceptation de manière à compléter les arguments avancés par Tennant ? Dans la littérature récente sur le déflationnisme, d'autres tentatives ont été proposées. Dans un article (Horsten et Leigh, 2017), consacré principalement à une étude des rapports entre principes de réflexion, clauses compositionnelles et théories décitationnelles, Leon Horsten et Graham Leigh abordent en quelques pages la question de la justification des

290. Ce qui avec les notations employées ici, se formalise justement par $A(\varphi) \rightarrow J(\varphi)$.

principes de réflexion. Ils font également appel aux travaux de Tyler Burge [291] sur la connaissance de soi pour justifier les schémas de réflexion sans faire usage de la notion de vérité. La justification proposée (*cf.* Horsten et Leigh (2017, section 6 p. 15-20)) reste toutefois à un niveau assez informel. À notre connaissance, la tentative d'analyse d'un processus réflexion la plus développée dans la littérature est celle de Cieśliński (2017) vers laquelle nous nous tournons à présent.

4.3.3.6 La théorie de la crédibilité de Cieśliński (2017)

Cieśliński (2017) consacre un ultime et long chapitre, point d'aboutissement de son ouvrage, à la présentation de sa propre solution du problème de la conservativité. En s'inspirant à la fois de Tennant (2002, 2010, 2005) et d'Horwich (2001, 1998b, 2010), il tente de proposer une solution uniforme commune au problème de la conservativité et au problème de la généralisation [292]. Face à ces deux problèmes, la stratégie déployée est la même. Elle consiste à montrer que certains énoncés qu'une théorie de la vérité déflationniste échoue à démontrer doivent néanmoins être acceptés par un agent rationnel ayant fait sienne une théorie d'arrière-plan. Dans l'ensemble, la tentative de solution avancée par Cieśliński (2017) est donc semblable à celles développées à la suite des réflexions de Tennant (2002). Il s'agira d'introduire un « processus de réflexion » censé pouvoir s'exercer indépendamment de toute notion substantielle de vérité et qui permettra de justifier l'adoption de divers énoncés indémontrables dans la théorie d'arrière-plan, au premier rang desquels divers principes de réflexion [293]. L'originalité du travail de Cieśliński (2017) consiste principalement à nos yeux à proposer un grand niveau de détails

291. Sont invoqués, plus précisément : Burge (2013, 1998, 1993, 2003).

292. Nous avons déjà évoqué en passant ce problème de la généralisation au cours de notre travail. Ce problème concerne la faiblesse déductive de certaines théories aléthiques déflationnistes, qui bien qu'elles permettent de prouver la collection infinie des instances de certaines propriétés, échouent à établir l'énoncé général, *i.e.* universellement quantifié, correspondant. Par exemple, rappelons qu'une théorie purement décitationnelle permet bien de dériver pour chaque énoncé φ une propriété du type $Vr(\varphi) \vee Vr(\neg\varphi)$ mais qu'elle ne permet pas d'établir l'énoncé général correspondant, à savoir $\forall x(En(x) \rightarrow (Vr(x) \vee Vr(neg(x))$. Ce problème avait déjà été noté par Tarski (1935), qui qualifiait une théorie de la vérité réduite aux seules **T**-équivalences de

> système hautement incomplet, auquel manqueraient les lois, de nature générale, les plus importantes et les plus féconde. (Tarski, 1935, p. 251)

C'est cependant Gupta (1993) qui le premier a mis en avant ce phénomène dans le cadre d'une critique du déflationnisme.

293. Petite nuance, pour Cieśliński (2017), on peut également employer ce processus de réflexion à partir d'une théorie purement décitationnelle de la vérité de manière à justifier l'adoption des clauses compositionnelles tarskiennes afin de surmonter le problème de la généralité.

dans l'analyse de ce processus et surtout à développer une véritable théorie axiomatique en vue de le formaliser.

Voici comment Cieśliński lui-même présente sa démarche :

> Le problème de la faiblesse déductive d'une théorie donnée peut être surmonté de deux manières différentes. Une méthode consiste à dériver les énoncés manquant dans une théorie plus riche. Par exemple, étant donné PA comme point de départ et étant donné que l'énoncé de la cohérence « $Con(PA)$ » est traité comme quelque chose devant être expliqué, on peut essayer de compléter l'arithmétique de Peano avec des axiomes supplémentaires —justifiés de manière indépendante— qui donneront ensuite « $Con(PA)$ » comme théorème. En d'autres termes, on peut tenter de *prouver* les énoncés manquants dans des théories plus riches, justifiées de manière indépendante.
>
> La seconde stratégie est celle qui sera mise en œuvre ici. [...] Étant donnée une théorie S d'arrière-plan, la tâche consiste à expliquer pourquoi nous *devrions* accepter divers énoncés, peut-être improuvables dans S elle-même. En choisissant cette seconde stratégie, nous ne tentons pas de prouver des énoncés qui sont indépendants de S. Ce qu'on essaye de faire à la place, c'est de démontrer que de tels énoncés *devraient être acceptés*. Ce qui est différent de les prouver.
>
> Dans le cadre formel présenté dans ce qui suit, cette approche correspondra à prouver non pas un énoncé indépendant φ lui-même mais la *crédibilité* [N.D.T *believability*] de φ. Dans cette optique, la thèse sera que quiconque accepte une théorie S devrait en fait accepter certains autres énoncés indémontrables dans S, en vertu de contraintes de rationalité. C'est-à-dire, en vertu du fait que notre conception de la crédibilité [N.D.T *conception of believability*] nous permet de dériver que ces énoncés supplémentaires sont crédibles. En résumé, l'explication proposée de nos engagements épistémiques se déroule comme suit. Si vous acceptez S, vous êtes alors tenus de considérer comme également crédibles d'autres énoncés supplémentaires qui ne sont pas impliqués par S elle-même.
>
> (Cieśliński, 2017, p. 253)

Remarquons que le processus de réflexion va s'appuyer chez Cieśliński (2017) sur un concept de « crédibilité [*believability*] » introduit par l'auteur lui-même, plutôt que sur des

notions d' « acceptabilité », d'« acceptation justifiée » ou d' « acceptation rationnelle » que nous avons rencontrées jusqu'ici. Mais, si la terminologie change, la démarche reste globalement la même.

Examinons à présent plus en détails, les étapes de sa construction. Comme Cieśliński lui-même le souligne, une étape préalable indispensable à l'analyse d'un processus de réflexion à la Tennant, est la clarification de ce que peut signifier précisément « accepter une théorie ». En s'appuyant sur les réflexions de Franzén (2004), Cieśliński (2017) isole quatre interprétations possibles de ce en quoi peut consister le fait d'accepter une théorie. Selon lui,

> « J'accepte PA » pourrait signifier :
>
> (a) J'accepte que tous les théorèmes de PA sont vrais.
>
> (b) J'accepte un certain type de principe de réflexion arithmétique pour PA (un principe uniforme [*i.e.* $RFN(PA)$] ou local [*i.e.* $Rfn(PA)$]).
>
> (c) Pour tout énoncé φ, si je croyais que φ possède une preuve dans PA et que je n'avais pas de raison indépendante de ne pas croire φ, alors je serais prêt à accepter φ. [...]
>
> (d) Pour toute formule $\varphi(x)$, si je croyais que (pour tout n, $\varphi(n)$ possède une preuve dans PA) et que je n'avais pas de raison indépendante de ne pas croire $\forall x\varphi(x)$, alors je serais prêt à accepter $\forall x\varphi(x)$.
>
> (Cieśliński, 2017, p. 242)

Ces interprétations ne peuvent cependant pas toutes convenir dans le cadre d'une défense du déflationnisme face à l'argument de la conservativité. La première (a) s'appuie sur une notion forte de vérité. Elle est donc inenvisageable pour le déflationniste. Les interprétations (b) et (d), quant à elles, permettent bien apparemment de surmonter le dilemme posé par l'argument, mais elles ne le font qu'en trivialisant totalement le problème : si, par définition, accepter PA c'est accepter $Rfn(PA)$ (interprétation (b)), alors toute personne acceptant PA...acceptera $Rfn(PA)$. Elles reviennent donc à purement et simplement postuler ce qu'il s'agit de justifier. Et c'est pourquoi Cieśliński (2017) les écarte également. En conséquence, Cieśliński (2017) retient finalement comme interprétation la définition (c) qu'il considère comme étant celle « présentant le plus grand défi » pour le déflationnisme. S'inspirant à nouveau de Franzén (2004), Cieśliński (2017, p.

246) identifie formellement cette manière d'accepter une théorie avec l'adoption de la règle d'inférence suivante :

$$\textbf{(R)}\ \frac{T' \vdash Thm_T(\varphi)}{\vdash \varphi}$$

où T' est une extension de T (peut-être T elle-même) et où Thm_T est un prédicat de prouvabilité dans T. Cette règle, que Cieśliński (2017) rebaptise règle de réflexion faible est semblable à la règle de Parikh, ou règle de réflexion locale que nous avons déjà évoquée (*cf.* page 370). Elle est connue pour être conservative sur PA. Insistons donc à nouveau ici sur ce point crucial : adopter une *règle* de réflexion telle que (R) donne une extension plus faible qu'adopter un *schéma/principe* de réflexion du type $Rfn(T)$ (local) ou $RFN(T)$ (uniforme). Dès lors, si l'on adopte cette interprétation de ce en quoi consiste accepter PA, on n'a pas, à première vue, de raison d'accepter un principe de réflexion :

> Puisque l'arithmétique de Peano augmentée de (R) n'est rien d'autre que PA [...], il reste encore à expliquer pourquoi quelqu'un qui accepte PA en ce sens devrait être engagé à accepter quelque énoncé que ce soit non prouvable dans PA.
> (Cieśliński, 2017, p. 246)

Autrement dit, comment expliquer qu'un agent ayant accepté PA au sens qui vient d'être exposé, est néanmoins justifié, voire tenu, d'accepter un principe de réflexion pour PA ? C'est ici qu'intervient le « processus de réflexion » dont Cieśliński va donner sa propre version.

Franzén (2004) affirmait qu'une personne acceptant PA, au sens où il adopterait la règle (R), tout en refusant d'adopter l'ensemble des implications de la forme « Si PA prouve φ alors φ », *i.e.* tout en refusant d'accepter un principe de réflexion donné sous la forme d'un ensemble d'énoncé implicatifs, ferait en réalite preuve d'irrationalité. La raison avancée par Franzén (2004) était que toute justification de la règle (R) vaut justification d'un principe de réflexion. Voici un passage illustrant bien ce type d'argumentation :

> Il est difficile d'envisager une quelconque justification pour considérer un théorème comme vrai sur cette base [294] qui n'implique pas d'accepter le prin-

294. N.D.T Franzén fait ici référence au fait d'accepter un théorème sur la seule base de l'information

> cipe de réflexion sous la forme d'un énoncé hypothétique plutôt que comme une simple règle d'inférence. Si nous ne sommes pas prêt à asserter l'énoncé hypothétique « si φ est prouvable dans T alors φ », sur quoi peut-on s'appuyer pour conclure φ en possédant uniquement l'information que φ est prouvable dans T (plutôt qu'une preuve dans T que nous avons examinée attentivement) ? (Franzén, 2004, p. 216) [295]

Toutefois, Cieśliński (2017) n'est *pas* convaincu par l'argumentation de Franzén (2004). Il considère qu'il peut exister des situations où se justifie l'adoption d'une *règle* faible de réflexion telle que (R) sans adopter l'ensemble des implications formant les *schémas* de réflexion tels que $Rfn(T)$ ou $RFN(T)$. Il écrit ainsi que :

> [...] en général, accepter la réflexion complète [N.D.T, c'est-à-dire un *schéma* de réflexion] demande une justification plus forte que pour adopter la règle de réflexion. Nous pouvons en effet adopter la règle pour des raisons qui sont trop faibles pour nous conduire à accepter une forme plus forte de réflexion. (Cieśliński, 2017, p. 249)

Aux yeux de Cieśliński (2017), il n'y a donc rien d'irrationnel en général à accepter (R) sans accepter Rfn ou RFN. Pour justifier l'adoption de principes de réflexion à partir de la simple acceptation de PA et pour répondre, ce faisant, au défi lancé aux déflationnistes par Ketland et Shapiro, Cieśliński (2017) va donc devoir emprunter une autre voie.

Cette voie va essentiellement consister à prendre au sérieux l'idée que lorsque nous acceptons une théorie, nous traitons les preuves formulées dans le cadre de cette théorie comme des raisons suffisamment probantes d'accepter leurs conclusions. Cieśliński (2017) déclare ainsi que

> la pratique consistant à accepter de nouveaux énoncés à partir de l'information qu'ils sont prouvables dans PA (même sans vérifier la preuve réelle), serait en effet irrationnelle à moins d'admettre que les preuves dans PA constituent de bonnes raisons d'accepter leurs conclusions. (Cieśliński, 2017, p. 251)

qu'il a été prouvé dans T mais sans examiner ou vérifier la preuve elle-même, ce que Cieśliński identifie avec l'adoption de la règle (R).

295. cité *in* Cieśliński (2017, p. 247).

Pour donner plus de corps à cette idée, Cieśliński (2017) introduit le concept de « crédibilité » d'un énoncé qu'il va s'efforcer d'analyser au point même d'en proposer une véritable théorie formalisée.

> [...] Supposons que « φ est crédible » signifie « il y a une raison probante d'accepter φ ». [...] En ces termes, on peut dire que c'est précisément la prouvabilité de φ qui le rend crédible.
>
> (Cieśliński, 2017, p. 251)

Plus précisément, dire que φ est crédible signifie pour Cieśliński que

> nous avons une raison d'accepter φ qui est *normalement* suffisamment probante, « normalement » signifiant ici « en l'absence d'autres raisons fortes d'accepter la négation de φ ». (Cieśliński, 2017, p. 251, italiques de l'auteur)

Comme Cieśliński le remarque lui même dans une note de bas de page, cette notion de crédibilité est taillée sur mesure pour convenir à l'interprétation (c) de ce en quoi consiste accepter PA[296] avec pour conséquence que

> la crédibilité de φ ne garantit pas automatiquement l'acceptation rationnelle de φ (Cieśliński, 2017, p. 251, note de bas de page)

C'est sur cette notion de crédibilité que Cieśliński (2017) va s'appuyer pour mettre au point sa propre version du « processus de réflexion ». Il poursuit ainsi :

> En fait, je propose de reconstruire le processus de réflexion en m'appuyant sur ce nouvel élément que je viens de mentionner. En l'occurrence, réfléchissant sur ma pratique mathématique, j'en viens à croire que :
>
> $(*)$ Pour tout ψ, si $PA \vdash \psi$, alors ψ est crédible.
>
> Refuser d'accepter (*) revient à envisager sérieusement la possibilité qu'il existe un ψ tel que $PA \vdash \psi$ mais qu'en même temps ψ ne soit pas crédible. En d'autres termes, si je refuse d'accepter (*), alors je ne suis pas convaincu que l'existence d'une preuve dans PA est (normalement) une raison probante de croire que ψ. Cependant, dans ma pratique je traite les preuves dans PA exactement comme de telles raisons. Dans une telle situation, ma position

296. *Cf.* les différentes significations possibles de « J'accepte PA » envisgées par Cieśliński (2017) que nous avons rappelées ci-dessus 4.3.3.6.

> est effectivement instable, dans la mesure où toute raison justifiant la continuation de ma pratique justifie également la crédibilité des théorèmes de PA.
>
> (Cieśliński, 2017, p. 251)

Le principe $(*)$ qui affirme la « crédibilité » des théorèmes de PA, plutôt que leur vérité, va donc être au cœur du processus de réflexion envisagé par Cieśliński (2017) :

Processus de réflexion selon Cieśliński

Voici comment Cieśliński lui-même distingue les étapes du processus de réflexion qu'il entend décrire :

> (1) Une personne P adopte la règle de réflexion faible : après avoir réalisé que φ est prouvable dans PA, elle est prête à accepter φ.
>
> (2) Réfléchissant sur sa pratique, P constate qu'elle traite les preuves dans PA comme des raisons qui sont normalement suffisamment bonnes pour accepter leurs conclusions (en résumé, P considère que les preuves dans PA rendent leurs conclusions crédibles [*believable*]).
>
> (3) Si une personne rationnelle possède une raison suffisamment bonne de croire φ (tout en n'ayant aucune bonne raison d'accepter la négation de φ), alors elle en vient à croire φ.
>
> Je considère ceci comme une explication partielle plausible de ce que l'on peut raisonnablement attendre d'un individu rationnel réfléchissant sur sa pratique mathématique. Cette pratique (c'est-à-dire l'adoption de la règle de réflexion faible) est fondée sur un énoncé hypothétique de crédibilité (voyez (2)) qui devrait être accepté sous peine de rendre la pratique de P irrationnelle. À cela s'ajoute une contrainte générale de rationalité, qui nous force à accepter des énoncés lorsque l'on a des raisons suffisantes de les accepter (voyez (3)). Il vaut la peine de noter que la notion de vérité n'entre nulle part en considération ici.
>
> (Cieśliński, 2017, p. 252)

L'étape (1) ci-dessus correspond à l'adoption de la règle (R), c'est-à-dire à l'acceptation de PA au sens qui a été retenu par Cieśliński (2017). Après « réflexion », l'étape (2) consiste à prendre conscience que nous traitons les preuves dans PA comme de bonnes

raisons pour accepter leurs conclusions, ce qui correspond à (l'adoption d') un principe tel que (∗). L'étape (3) aura aussi son importance. Mais pour voir plus précisément comment ces étapes doivent permettre de justifier, *in fine*, l'adoption d'un principe tel que Rfn ou RFN ou l'adoption d'un énoncé de cohérence tel que $Con(PA)$, il faut analyser plus en profondeur la notion de crédibilité avancée par Cieśliński. C'est à cette tâche que Cieśliński (2017) s'attelle en introduisant un nouveau prédicat unaire $B(x)$, dont l'interprétation attendue sera « x est crédible [*beliveable*] », et en proposant une véritable théorie axiomatisée pour en caractériser la signification. Nous exposons donc à présent les principaux éléments techniques de Cieśliński (2017, chapitre 13). Nous reprenons les formulations de l'auteur et nous contentons d'énoncer les résultats sans reproduire les preuves qui sont parfois longues et relativement complexes [297].

Pour formaliser la notion de crédibilité [*Believability*] qu'il a introduite, Cieśliński (2017) donne les axiomes et les règles d'inférence suivants :

Définition (Cieśliński (2017), p. 254)**.** Soit K une extension axiomatisable de PA formulée dans le langage $\mathcal{L}_K$ (qui peut être plus riche que $\mathcal{L}_{PA}$). Soit $\mathcal{L}_{K,B}$ l'extension de $\mathcal{L}_K$ obtenue par l'ajout d'un nouveau prédicat unaire « B ». Soit KB la théorie K *formulée dans le langage* $\mathcal{L}_{K,B}$ [298]

— On notera $Bel(K)^-$ la théorie formulée dans le langages $\mathcal{L}_{K,B}$ qui étend KB au moyen des axiomes suivants :

(A_1) $\forall \psi \in \mathcal{L}_{K,B}\ [Thm_{KB}(\psi) \rightarrow B(\psi)]$

(A_2) $\forall \varphi, \psi \in \mathcal{L}_{K,B}\ [B(\varphi) \wedge B(\varphi \rightarrow \psi) \rightarrow B(\psi)]$

En outre, $Bel(K)^-$ contient les deux nouvelles règles d'inférence suivantes :

$$\textbf{NEC}\ \frac{\vdash \phi}{\vdash B(\phi)} \qquad\qquad \frac{\vdash \forall x B(\phi(x))}{\vdash B(\forall x \phi(x))}\ \textbf{GEN}$$

— On notera $Bel^{Con}(K)^-$ la théorie $Bel(K)^-$ complétée au moyen de l'axiome de consistance suivant :

(A_3) $\forall \psi \in \mathcal{L}_{K,B}\ \neg B(\psi \wedge \neg \psi)$

297. Pour plus de détails nous renvoyons évidemment le lecteur à Cieśliński (2017, chapitre 13, p. 254-266).

298. Autrement dit, les axiomes et règles de la logique du premier ordre sont étendus à l'ensemble des formules du langage $\mathcal{L}_{K,B}$, *i.e.* y compris celles qui contiennent le nouveau prédicat « B ». En revanche les axiomes non-logiques de K et de KB sont rigoureusement les mêmes (et ne contiennent donc pas le nouveau prédicat).

— Enfin, on notera $Bel(K)$ et $Bel^{Con}(K)$ les théories qui sont exactement comme $Bel(K)^-$ et $Bel^{Con}(K)^-$ à ceci près qu'elles contiennent tous les axiomes d'induction pour les formules de $\mathcal{L}_{K,B}$.

L'axiome (A_1) est une simple formalisation du principe $(*)$ affirmant que les théorèmes de KB[299] sont crédibles. L'axiome (A_2) énonce une propriété de clôture de la crédibilité et ne semble pas poser de problème particulier : si j'ai une bonne raison de croire une implication et une bonne raison de croire son antécédent, alors j'ai une bonne raison d'en croire le conséquent. Les nouvelles règles d'inférence introduites dans $Bel(K)$ appellent plus de commentaires. Pour Cieśliński,

> La validité intuitive de **NEC** [...] semble évidente et non sujette à controverse. Une fois fournie une preuve de φ dans $Bel(K)$, c'est simplement cette preuve elle-même qui est considérée comme une raison probante d'accepter φ, rendant de fait φ crédible. (Cieśliński, 2017, p. 255)

La règle **GEN** est quant à elle cruciale d'après Cieśliński (2017). C'est elle qui va permettre de tirer des conséquences fortes de $Bel(K)$. Pour la justifier, Cieśliński souligne que

> Pour pouvoir appliquer la règle, il nous faut une preuve dans $Bel(K)$ de l'énoncé général « $\forall n B(\psi(n))$ », et cette preuve elle-même fournit une raison probante uniforme de croire que toutes les instances de ψ sont crédibles. L'intuition est alors qu'une raison probante de considérer que toutes les instances sont crédibles constitue également une raison probante de croire l'énoncé général. (Cieśliński, 2017, p. 255)

L'axiome (A_3), qui affirme en substance qu'une contradiction n'est jamais crédible, est plus problématique. Cieśliński (2017) prend soin de le laisser de côté et de ne pas en faire un axiome de sa théorie minimale de la crédibilité $Bel(K)$. C'est un axiome qu'on peut choisir ou non d'intégrer à notre conception de la crédibilité, selon les intuitions que l'on peut entretenir à l'égard de cette notion. Au total, pour une théorie d'arrière-plan K qu'un agent rationnel accepte, $Bel(K)$ constitue une théorie axiomatique de la « crédibilité de K » proposée par Cieśliński. Elle pourra être mobilisée pour expliciter le processus de réflexion dans lequel un agent ayant accepté PA peut s'engager. Avant d'en

299. *i.e.* de la théorie de base K qu'on a initialement acceptée, simplement augmentée d'axiomes logiques pour les énoncés contenant le nouveau prédicat $B(x)$.

venir aux divers résultats techniques à propos de cette théorie, il faut souligner, comme Cieśliński lui-même le remarque, que

> l'interprétation attendue de « $B(x)$ » se manifeste non pas tant dans la mécanique de preuve de $Bel(K)$ [...] mais plutôt dans l'opération —*non formalisée dans la théorie*— qui nous mène de l'acceptation de « $B(\psi)$ » à l'acceptation de ψ. (Cieśliński, 2017, p. 256, nous soulignons)

Ce passage de $B(\psi)$ à ψ correspond en fait à l'étape (3) du processus de réflexion esquissé ci-dessus (*cf.* page 409). Il est important de bien voir qu'il ne fait pas partie de $Bel(K)$ elle-même.

Venons-en à présent aux propriétés techniques de $Bel(K)$ obtenues par Cieśliński (2017). Il montre tout d'abord que $Bel(K)$ est une théorie « sûre » au sens où elle est interprétable dans le modèle standard de l'arithmétique. Plus précisément, Cieśliński introduit tout d'abord la définition suivante :

Définition (Cieśliński (2017), p. 256)**.** $Int_{Bel(K)} = \{\psi \in \mathcal{L}_{K,B} / Bel(K) \vdash B(\psi)\}$

$Int_{Bel(K)}$ (l'« intérieur » de $Bel(K)$) désigne l'ensemble des énoncés dont $Bel(K)$ prouve qu'ils sont crédibles. Notez que, par la présence de la règle de nécessitation **NEC** contenue dans $Bel(K)$, $Bel(K) \vdash \phi$ implique $Bel(K) \vdash B(\phi)$ ce qui a pour conséquence que $Bel(K)^{\vdash} \subset Int_{Bel(K)}$, *i.e.* toutes les conséquences déductives de $Bel(K)$ sont contenues dans l'intérieur de $Bel(K)$. L'inclusion inverse $Int_{Bel(K)} \subset Bel(K)^{\vdash}$ n'est en général pas vérifiée. En particulier du fait de la règle de généralisation **GEN** rien n'empêche *a priori* qu'il existe des énoncés ψ tels que $Bel(K) \vdash B(\psi)$ alors que $Bel(K) \nvdash \psi$. Ce sera notamment le cas lorsque $K = PA$ et que ψ est un schéma de réflexion pour PA : nous allons voir, par exemple que $Bel(PA)$ prouve qu'un certain schéma de réflexion équivalent à $RFN(PA)$ est « crédible » —*i.e.* $Bel(PA) \vdash B(\forall x[Thm_{PA}(\varphi(x)) \rightarrow \varphi(x)])$ pour tout énoncé $\varphi \in \mathcal{L}_{PA}$— mais ne prouve pas ce schéma de réflexion lui-même —*i.e.* $Bel(PA) \nvdash \forall x[Thm_{PA}(\varphi(x)) \rightarrow \varphi(x)]$.

Concernant $Bel(K)$, le premier théorème obtenu par Cieśliński (2017) est le suivant :

Théorème (Cieśliński (2017), p. 257)**.** Soit K une théorie exprimée dans le langage $\mathcal{L}_K$ étendant $\mathcal{L}_{PA}$ mais ne contenant pas le prédicat « $B(x)$ ». Si $\mathbb{N}$ (le modèle standard de l'arithmétique) peut être étendu à un modèle $\mathbb{N}^*$ de K, alors $\mathbb{N}^*$ peut être étendu à un modèle de $Int_{Bel(K)}$.

A fortiori, $\mathbb{N}^*$ peut donc être étendu à un modèle de $Bel(K)$. Ceci établit que si l'on part d'une théorie K d'arrière-plan interprétable dans le modèle standard de l'arithmétique, alors la théorie étendue $K \cup Bel(K)$ sera elle aussi interprétable dans ce modèle. Elle sera donc cohérente (et même ω-cohérente). Par conséquent, on peut, sans crainte de mauvaises surprises, s'appuyer sur cette théorie pour formaliser un processus de réflexion sur l'arithmétique de Peano. Les résultats qui suivent viennent précisément jouer ce rôle. Tout d'abord Cieśliński formule la définition suivante :

Définition (Cieśliński (2017), p. 25 ?)**.** En partant de l'arithmétique de Peano PA exprimée dans $\mathcal{L}_{PA}$, on enrichit le langage au moyen d'un nouveau prédicat unaire « $B(x)$ » et on définit la suite de théories suivantes :

— $S_0 = PAB$ [300]

— $S_{n+1} = S_n \cup \{\forall x[Thm_{S_n}(\varphi(x)) \rightarrow \varphi(x)]/\varphi(x) \in \mathcal{L}_{PA,B}\}$

Partant de l'arithmétique de Peano augmentée d'un nouveau prédicat de crédibilité, on a donc défini une suite croissante de théories ($S_n \subset S_{n+1}$) exprimées dans le langage $\mathcal{L}_{PA} \cup B(x)$, où à chaque étape S_{n+1} est obtenue en adjoignant à S_n un schéma de réflexion

$$\forall x\Big[Thm_{S_n}(\varphi(x)) \rightarrow \varphi(x)\Big]$$

formulé pour toutes les formules $\varphi(x)$ du langage $\mathcal{L}_{PA} \cup B(x)$.

Cieśliński (2017) démontre ensuite les deux résultats suivants :

Théorème (Cieśliński (2017), p. 25 ?)**.** Pour tout entier naturel n,

$$Bel(PA) \vdash \forall \varphi \in \mathcal{L}_{PA,B}\ (Thm_{S_n}(\varphi) \rightarrow B(\varphi))$$

Théorème (Cieśliński (2017), p. 25 ?)**.** Pour tout entier naturel n, pour tout énoncé $\varphi \in \mathcal{L}_{PA,B}$,

$$Bel(PA) \vdash B\Big(\forall x[Thm_{S_n}(\varphi(x)) \rightarrow \varphi(x)]\Big)$$

Le premier résultat établit que la théorie de la crédibilité $Bel(PA)$ permet de prouver pour chaque théorie S_n un énoncé équivalent à $(*)$, autrement dit prouver que chaque théorie S_n est crédible. Le second théorème ci-dessus énonce que le schéma de réflexion :

300. S_0 est donc semblable à PA à ceci près que les axiomes et règles logiques sont étendus aux énoncés contenant le nouveau prédicat $B(x)$.

$$\forall x[Thm_{S_n}(\varphi(x)) \rightarrow \varphi(x)]$$

est « crédible » pour chaque théorie S_n. C'est, nous semble-t-il, le résultat central pour la mise au jour d'un processus de réflexion permettant de répondre au problème de la conservativité. En effet, $Bel(PA)$ prouve en particulier

$$B(\forall x[Thm_{PA}(\varphi(x)) \rightarrow \varphi(x)]),$$

autrement dit, $Bel(PA)$ établit la crédibilité de ce schéma de réflexion pour PA. De plus, dans PA, on peut montrer pour chaque entier n, que les schéma de réflexion ci-dessus sont équivalents à $RFN(S_n)$. Ainsi, la théorie axiomatisant la notion de crédibilité permet de dériver un énoncé exprimant la crédibilité d'un schéma de réflexion uniforme pour PA et également pour toute une série de théories $(S_n)_{n\in\mathbb{N}}$ étendant PA.

Si l'on reprend les étapes du processus de réflexion tracé par Cieśliński, on aurait donc la séquence suivante :

(1) À la première étape une personne P accepte PA, en adoptant la règle de réflexion (R) (qui est conservative sur PA).

(2) Réfléchissant sur sa pratique mathématique, P réalise qu'elle traite les preuves dans PA comme des raisons probantes d'accepter leurs conclusions. Elle en vient alors à adopter un énoncé tel que $(*)$ ou (A_1) qui énonce que les théorèmes de PA sont « crédibles ». Toutefois, cette étape ne s'arrête pas là. P réfléchit en outre aux propriétés de la « crédibilité » des théorèmes qu'elle adopte, ou pour le dire autrement, P réfléchit aux propriétés des raisons probantes qui président, dans sa pratique mathématique, à son acceptation de tel ou tel énoncé comme théorème. Elle en vient alors à accepter, outre (A_1), certains énoncés formulant des propriétés de la crédibilité. Elle adopte ainsi des énoncés comme (A_2) et des règles telles que **NEC** et **GEN**. Bref, elle adopte quelque chose comme une théorie de la crédibilité formalisable par $Bel(PA)$. À l'aide de cette théorie P peut établir, c'est-à-dire prouver, que certains énoncés, en particulier certains principes de réflexion indémontrables dans PA, sont néanmoins crédibles.

(3) Arrive alors la troisième étape : une fois munie d'une preuve de la crédibilité de certains énoncés, P considère qu'elle possède une raison probante d'accepter ces énoncés. Et en l'absence de raison probante de ne pas accepter ces énoncés, P agit donc rationnellement en acceptant ces énoncés.

Voici comment Cieśliński (2017) lui-même résume la démarche de son processus de réflexion :

> Dans notre pratique, nous traitons les preuves dans [notre théorie T] comme des raisons valables d'accepter leurs conclusions. En réfléchissant sur cette pratique, nous reconnaissons tout d'abord que c'est effectivement ce que nous faisons. En second lieu, nous prenons conscience qu'une telle pratique serait irrationnelle sans la croyance sous-jacente que les théorèmes de [T] sont crédibles (c'est-à-dire, sans la croyance que les preuves dans [T] sont normalement suffisamment bonnes pour nous faire accepter leurs conclusions).
>
> À l'étape suivante, nous tentons de caractériser notre notion de crédibilité au moyen de certains axiomes et règles simples et fondamentaux. En conséquence, nous finissons par déclarer crédibles certains énoncés supplémentaires formulés dans le langage de [T] (énoncés qui ne sont pas prouvables dans [T] elle-même). Notre acceptation initiale de [T], associée avec certaines convictions fondamentales concernant la crédibilité, nous conduit à reconnaître qu'en fait nous avons une raison probante d'accepter, par exemple, des clauses compositionnelles pour la vérité ou des principes de réflexion. Devant une telle raison probante, nous agissons rationnellement quand, finalement, nous les acceptons. (Cieśliński, 2017, p. 266-267)

Autrement dit, ce processus de réflexion et ce passage par une théorie formalisant les propriétés de la crédibilité, permet de justifier l'acceptation de certains énoncés qui n'ont pas *stricto sensu* été démontrés mais dont la crédibilité a en revanche été établie. Dans ce processus, la notion de vérité n'a aucune part. Tout repose sur une analyse de la manière dont nous traitons certaines preuves comme des raisons probantes d'accepter leurs conclusions et sur la notion de crédibilité qui en découle, associée à certaines contraintes de rationalité. La mise au point d'un tel processus de réflexion permettant de justifier l'acceptation de principes de réflexion tels que $(Rfn(PA))$ ou $RFN(PA)$ est sans conteste une grande réussite pour les partisans du déflationnisme. Elle semble permettre de résoudre le dilemme posé par Shapiro et Ketland. En effet, muni d'un tel outil, le déflationniste peut proposer, en réponse à l'argument de la conservativité, de rénoncer à l'exigence de réflexivité pesant sur une théorie adéquate de la vérité [301]. Puisque l'on

301. Ce qui correspond à écarter la seconde prémisse de l'argument. Voyez notre rappel du squelette de l'argument en début de chapitre, page 261.

peut justifier autrement l'acceptation de $Rfn(PA)$, $RFN(PA)$ et $Con(PA)$, il n'est plus nécessaire d'exiger que notre extension aléthique permette de prouver de tels principes ou énoncés. Le déflationniste peut alors se cantonner à une théorie « modeste » de la vérité, par exemple à une théorie conservative [302].

Toutefois, la construction proposée par Cieśliński (2017) ne nous semble pas clore définitivement ce débat. Elle soulève en effet un certain nombre de difficultés dont certaines sont envisagées par Cieśliński lui-même dans les dernières pages de son ouvrage. Nous ne prétendrons pas en proposer ici une analyse approfondie mais nous nous contenterons d'évoquer des pistes de recherche pour un travail futur.

La théorie de la crédibilité développée par Cieśliński (2017) est formellement impeccable et peut sembler à première vue plausible. Elle soulève néanmoins divers problèmes d'interprétation. Cieśliński lui-même le reconnaît et paraît parfois osciller entre diverses lectures plus ou moins fortes de sa notion de crédibilité. Il faut d'abord bien voir que la nature de la justification obtenue au bout d'un processus de réflexion appuyé sur $Bel(K)$ est très différente de la preuve sémantique obtenue au moyen des clauses tarskiennes pour la vérité. Scrupules déflationnistes mis à part, rien d'absurde à première vue a affirmer que si une théorie est vraie alors ses théorèmes le sont également, ou bien encore à affirmer que si une théorie est vraie alors elle sera cohérente. En revanche, mon attitude épistémique envers une théorie, que ce soit croyance, acceptation, rejet ou autre, n'a pas

302. Même si cet autre aspect du travail de Cieśliński (2017) ne nous a pas retenus ici, signalons toute de même que Cieśliński met également sa théorie de la crédibilité au service d'une réponse au problème de la généralisation. Pour cela il s'appuie sur les résultats suivants, inspirés de Halbach (2001a, 2009) et Horsten et Leigh (2017) :

Théorème (Cieśliński (2017), p. 264)**.** $Bel(TB^-) \vdash B(CT^-)$

et

Théorème (Cieśliński (2017), p. 266)**.** $Bel(TFB^-) \vdash B(KF^-)$

où TB^- désigne une théorie purement décitationnelle typée de où CT^- désigne une théorie compositionnelle à la Tarski, tandis que $Bel(TFB^-) \vdash B(KF^-)$ exprime un résultat similaire dans une cadre non-typé. La démarche est alors la suivante :

(1) Le déflationniste adopte une théorie « modeste » purement décitationnelle comme TB. Cette théorie échoue à prouver certaines généralisations.

(2) Le déflationniste engage un processus de réflexion qui débouche sur une preuve de la crédibilité de principes sémantique plus forts, *e.g.* $B(CT^-)$. Ceci lui donne une justification pour accepter de tels principes.

(3) Grâce à ces nouveaux principes, le déflationniste peut établir les généralisations qui faisaient défaut à l'étape (1).

Au total, le processus de réflexion permet donc de surmonter également le problème de la généralisation.

d'impact *a priori* sur la question objective de savoir si cette théorie permet de dériver une contradiction ou si ses théorèmes sont vrais. Une fois acceptée une théorie T, quelle est donc la valeur justificative d'une dérivation dans $Bel(T)$ d'une formule établissant la crédibilité d'un énoncé φ ?

Nous avons vu qu'il existe des modes d'acceptation d'une théorie qui ne semblent pas sanctionner le processus de réflexion élaboré par Cieśliński (*cf.* 4.3.3.2). Bien plus, Cieśliński (2017) aborde des interprétations plus ou moins contraignantes de sa notion de crédibilité. Ainsi, lorsqu'il critique les propositions de Franzén (2004), Cieśliński (2017) affirme qu'il n'y a rien d'irrationnel pour un agent à accepter une théorie T (et à accepter la règle de réflexion faible (R) pour cette théorie) sans accepter des formes plus fortes de réflexion telles que $Rfn(T)$. Il donne même des exemples de situations dans lesquelles une telle attitude serait recommandée. Pour autant, lorsqu'il développe sa formalisation de la crédibilité, Cieśliński affirme qu'il serait irrationnel d'accepter PA sans accepter un principe tel que $(*)$, *i.e.* (A_1), puis évoque des contraintes de rationalité formalisées par $Bel(PA)$ pour dériver $B(Rfn(PA))$, avant de conclure qu'un agent ayant prouvé $B(Rfn(PA))$ agit rationnellement lorsque, en l'absence d'autres raisons probantes de douter de $Rfn(PA)$, il finit par accepter $Rfn(PA)$. Dès lors, on peut s'interroger : un agent qui accepte PA sans accepter $Rfn(PA)$ est-il ou n'est-il pas irrationnel ? Et, s'il n'est pas irrationnel, à quel point la dérivation de $B(Rfn(PA))$ est-elle contraignante ? À quel point nous oblige-t-elle à accepter $Rfn(PA)$?

De même, on peut se demander à quelles conditions, ou sur quelles théories le processus de réflexion évoqué par Cieśliński (2017) peut ou doit s'exercer. Doit-il s'exercer sur toute théorie que je pourrais accepter, quel que soit par ailleurs son contenu ? Sans doute que non. Par exemple, en l'état actuel de nos connaissances scientifiques il est manifestement rationnel d'accepter la théorie générale de la relativité. De même, il est manifestement rationnel d'accepter la théorie de la physique quantique. Pourtant, ces deux théories sont connues pour être mutuellement incompatibles. Sur laquelle de ces deux théories puis-je exercer un processus de réflexion ? Et, si l'une de ces deux théories que j'ai acceptées, disons $T_{E=mc^2}$, s'avérait fausse, faudrait-il en conclure que les axiomes de la théorie de la crédibilité $Bel(T_{E=mc^2})$ sont faux, ou plutôt que je n'aurais pas dû les appliquer à $T_{E=mc^2}$? Plus généralement, que faire si notre théorie de la crédibilité nous amène à déclarer crédibles deux énoncés contradictoires ?

Cieśliński (2017, p. 267-269) évoque ce problème et ébauche plusieurs scénarios

possibles. Il déclare ainsi,

> La question clef est de savoir comment la crédibilité doit être interprétée. Qu'est censé signifier, en termes intuitifs, $B(\varphi)$? Si cela signifie simplement que nous avons une raison probante d'accepter φ, bien que, dans le même temps, existe la possibilité d'une autre raison probante de rejeter φ, alors la dernière étape du raisonnement réflexif (celle où on accepte φ) ne semble toujours pas être justifiée. (Cieśliński, 2017, p. 267)

L'étape évoquée ici par Cieśliński est l'étape (3) du processus de réflexion [303], celle qui consiste à passer d'une preuve de $B(\varphi)$ à l'acceptation de φ lui-même. Cieśliński (2017) prend bien soin de préciser que cette étape n'appartient pas à $Bel(T)$ elle-même, *i.e.* qu'elle n'est pas formalisée à l'intérieur de la théorie axiomatisée de la crédibilité. Et pour cause, Cieśliński (2017) indique que ce passage de $B(\varphi)$ à φ n'est pas valable en toute généralité. Il n'est valable que « normalement » ou « en l'absence d'autres raisons probantes de rejeter φ ».

Quelques lignes plus haut, Cieśliński écrit également qu'

> on pourrait déclarer que
>
> (•) Pour tout énoncé φ, si je savais que φ est crédible et que je n'avais pas de raison indépendante de douter de φ, alors je devrais être prêt à accepter φ.
>
> Dans un tel cas, le simple fait de prouver $B(\varphi)$ ne suffit pas à garantir le passage qui mène à l'acceptation finale de φ. Il nous faudrait également l'information concernant l'absence de toute « raison indépendante de douter de φ ». (Cieśliński, 2017, p. 267)

On peut donc se demander quand la dernière étape du processus de réflexion, *i.e.* le passage de $B(\varphi)$ à φ, est réellement possible. Quand ce passage est-il « sûr » ? Autrement dit, quand est-ce que le processus de réflexion débouche non seulement sur une preuve $B(\varphi)$ mais encore sur l'acceptation (légitime) de φ ? Autrement dit encore, quand est-ce que ce processus nous fournit une justification *bona fide* d'un énoncé φ ? À quelles conditions sur φ ou sur la théorie d'arrière-plan que l'on a acceptée au départ ?

Face à ce type de problèmes, Cieśliński (2017) évoque différentes solutions. Ajouter à $Bel(T)$ un axiome de cohérence tel que (A_3) revient peu ou prou à postuler ce qu'il

303. *Cf.* page 409.

s'agit de justifier et Cieśliński (2017, p. 268) considère cette solution plus « comme un vol que comme du travail honnête ». Plus loin, il écrit

> Une solution modeste consisterait à restreindre[304] le champ d'applications de la théorie de la crédibilité. Avec cette approche, étant donné $B(\varphi)$, l'agent réflexif devrait rationnellement accepter φ tant qu'il n'est pas capable de dériver $B(\neg\varphi)$ dans une théorie considérée. (Cieśliński, 2017, p. 269)

Un autre scénario encore envisagé par Cieśliński (2017) consisterait à doter la théorie de la crédibilité d'une logique paraconsistante de manière à éviter le phénomène de l'explosion logique lorsqu'une contradiction est déclarée crédible. Toutefois, les remarques formulées par Cieśliński (2017) en ce sens restent pour la plupart à l'état programmatique. Bref, on le voit, le « processus de réflexion » n'a pas encore révélé tous ses mystères. Tenter de le clarifier de manière encore plus précise constituera sans doute une voie de recherche fertile pour les années à venir.

304. Mais selon quels critères précis ?

4.4 Annexes

4.4.1 Défense et illustration du déflationnisme en matière de mathématiques transfinies

En Attendant Gödel

Dans son bureau à l'Université Hilbert est très occupé. La tête entre les mains, il semble extrêmement soucieux et concentré. De temps à autre, il ne cesse de griffonner des formules sur des papiers disposés sur son bureau, sans parvenir à trouver la solution à ses problèmes. Tout à coup, on frappe à la porte. Un distingué collègue, appelons-le Kantor, entre alors.

Kantor : Mon cher Hilbert, vous m'avez l'air bien soucieux. On m'apprend que vous auriez été chassé du paradis ? J'entends que vous ne croiriez pas aux éléments transfinis et que vous ne voudriez pas accepter des explications qui s'appuieraient sur ce genre d'hypothèses.

Hilbert, peut-être légèrement agacé d'être ainsi interrompu dans son travail, reste plongé dans ses papiers.

Kantor : Mon cher collègue, sachez que moi aussi je partage totalement cette aversion pour le notions infinies abstraites, qu'on ne peut observer avec certitude de ses propres yeux et manipuler concrètement. D'ailleurs, il faut toujours se méfier de ces hôtels qui ne sont jamais complets et où il reste toujours des chambres libres en toute saison. C'est mauvais signe. Quand je voyage pour mes conférences, je tache toujours de faire mes réservations assez à l'avance pour éviter ce genre d'établissements. Mais, voyez-vous, j'ai la solution a vos problèmes.

À ces mots, Hilbert relève la tête, les yeux écarquillés par l'intérêt.

Hilbert : Que dites-vous là ! Est-ce que... Est-il possible que vous soyez parvenus à...

Kantor : *(Très agité et parcourant la pièce de part en part, faisant de grands gestes)* Très cher ami, la clef de tout ceci est de prendre conscience que les notions transfinies ne sont en fait que des outils expressifs ! En réalité elles n'ont pas de pouvoir explicatif, ou de contenu substantiel, qui pourrait mettre en danger la sûreté de nos recherches et réclamerait une analyse philosophique profonde. Tout comme vous, je

ne crois qu'en l'existence des notions finies, disons les nombres entiers, considérés comme des suites de marques sur le papier. Seuls ces objets concrets sont dotés de contenu et peuvent jouer un rôle explicatif dans nos constructions théoriques. Malgré cela, le langage de l'arithmétique est trop pauvre et expressivement étriqué ! Il faut l'enrichir de nouveaux instruments expressifs et c'est exactement dans cet esprit que je propose d'introduire ce qu'on qualifie avec plus ou moins de bonheur de notions transfinies. Avec mes collègues Sermelo et Frinkel, nous travaillons d'ailleurs en ce moment même sur les possibilités de proposer des extensions de ce type en se limitant à une axiomatisation très simple. Et, regardez le résultat : à présent je peux formuler des généralisations qu'il m'était impossible, ou peu commode, d'exprimer avant l'introduction de ces nouveaux outils. Ainsi, au lieu de fastidieusement donner une liste des entiers possédant telle ou telle propriété exprimable dans le langage de l'arithmétique (finitiste), je peux désormais parler de l' « ensemble » de ce ces entiers. Ici, le terme d' « ensemble » n'est évidemment qu'une façon de parler, un simple instrument expressif. Bien plus, si je voulais désigner les entiers vérifiant toute une série potentiellement infinie de propriétés, je serai bien en peine de le faire à l'aide du seul langage de base. Mais grâce à mes nouveaux moyens expressifs, je puis prendre l'« intersection » des « ensembles » caractérisés par ces propriétés. Et, ça ne s'arrête pas là, je peux réitérer ce processus et prendre des intersections de réunions, des réunions de réunions, *etc.* Voyez la richesse expressive nouvelle dont nous pouvons disposer ! J'oserai dire qu'elle est indispensable !

Hilbert : *(Songeur)*
Humm, en somme, vous pensez à un emploi purement instrumental des notions...

Kantor : Précisément, mon cher confrère !

Kantor s'approche du bureau et aperçoit les notes et les calculs de Hilbert éparpillés dessus.

Kantor : Ah ! Diantre ! Vous semblez troublé de ce que ma mathématique enrichie de ces nouvelles notions transfinies n'est peut-être pas conservative sur les mathématiques finitistes. Mais, cher camarade, vous avez bien tort ! Vos inquiétudes sont mal placées et dues uniquement à une mauvaise compréhension du caractère purement expressif des notions nouvellement introduites. En me dotant d'outils purement expressifs, je puis à présent faire des généralisations que je ne pouvais pas réaliser auparavant.

Et, ces généralisations peuvent avoir un impact sur le domaine des mathématiques finitistes. Il n'est pas surprenant ni controversé que l'augmentation du pouvoir expressif de mon langage me permette de formuler et de contracter des engagements nouveaux concernant le domaine des mathématiques contentuelles. Mais cela ne signifie nullement que les notions transfinies nouvellement introduites aient un quelconque contenu réel ni qu'elles jouent le moindre rôle explicatif au sein de nos théories. Je vous les répète : ce sont de simples outils expressifs. Allons, soyez rassuré ! Faites comme moi, adoptez une attitude raisonnablement et lucidement déflationniste en matière de mathématiques idéales et ne vous laissez plus ébranler par ces faux problèmes de conservativité.

À cet instant, on frappe à la porte. Un jeune étudiant polonais fait son entrée. Appelons-le Tarsky.

Tarsky : Cher maîtres, je vous prie de m'excuser de vous déranger ainsi. Mais je brûle de vous avertir : en m'appuyant sur les travaux de Monsieur le Professeur Kantor, j'ai pu obtenir un merveilleux résultat. Je suis parvenu à montrer comment la vérité peut-être étudiée à partir de la théorie des classes. Dans certains cas et modulo certaines hypothèses sur le langage employé, on obtient même une définition explicite de ce prédicat. Dans d'autres cas, une telle définition est impossible[305] mais on peut néanmoins employer une axiomatisation qui satisfera un critère d'adéquation que j'ai moi-même introduit. Bien sûr, ces « métathéories » de la vérité, si vous voulez bien me passer l'expression, seront bien souvent non conservatives. Mais, il est clair, ne serait-ce qu'en vertu du statut des outils que j'ai employés pour définir ou axiomatiser la vérité et qui sont tout à fait ceux dont parlait Monsieur le Professeur Kantor, il est clair, dis-je, que ceci n'est pas un problème. Nous savons bien que la vérité, à l'instar des notions dont parlait Monsieur le Professeur Kantor, est un outil purement expressif, indispensable seulement pour pouvoir augmenter le pouvoir expressif de notre langage. Comme le soulignait Monsieur le Professeur Kantor, il n'est d'ailleurs pas étonnant que l'emploi d'un tel outil puisse nous permettre de formuler des généralisations qui iront au delà de ce que nous pouvions dire au moyen du seul langage de mathématiques finitistes. N'est-ce pas formidable !

Hilbert, reprenant sa tête entre ses mains et le travail laissé inachevé sur son bureau, pousse alors un lourd et profond soupir.

305. À l'époque de ce dialogue, Tarsky n'avait pas encore écrit le Postscript à son travail.

Rideau

Ce que l'histoire ne nous dit pas, c'est si ce soupir poussé par Hilbert était un soupir de soulagement d'être enfin libéré de ses angoisses métaphysiques au sujet des mathématiques inifinitaires, ou de consternation face à la radicale et insurmontable incompréhension qui le séparait de ses collègues.

Conclusion

Au commencement de ce travail, nous avons demandé : « qu'est-ce que la vérité ? ». Nous l'avons fait essentiellement pour introduire et souligner l'originalité de la position déflationniste face à cette question. L'objectif de notre étude n'était pas en effet de proposer une réponse définitive à cette question qui occupe les philosophes depuis des millénaires, mais plus modestement d'évaluer la cohérence et la solidité du déflationnisme aléthique contemporain en le confrontant à certains arguments avancés contre ce type de conceptions philosophiques touchant la vérité. Au moment de conclure, nous rappelons les principaux points que notre travail a permis d'établir et nous en esquissons quelques prolongements possibles.

Une première étape de notre travail a consisté à exposer et clarifier la nature des positions déflationnistes actuelles. Ceci fut fait dans notre premier chapitre, où à travers une brève étude historique nous avons pu mettre en lumière les divers thèmes qui composent le déflationnisme contemporain en matière de vérité. Nous avons vu qu'à la suite des travaux de Quine, les principaux auteurs déflationnistes actuels ne veulent voir dans le prédicat de vérité qu'un outil de décitation purement expressif, indispensable —et donc inéliminable— pour formuler certaines généralisations mais dénué de tout contenu explicatif propre. Parallèlement, ces auteurs attribuent un rôle central aux **T**-équivalences en tant que théorie de la vérité : la collection infinie de ces équivalences est censée fournir une analyse exhaustive du concept de vérité. À elles seules, elles doivent permettre de rendre compte de tous nos usages légitimes du prédicat « vrai », d'expliquer complètement notre compréhension de ce concept et d'analyser entièrement sa signification. Une fois ces éclaircissements apportés, notre travail s'est poursuivi en deux parties distinctes que l'on peut voir comme deux tentatives complémentaires de fournir un cadre méthodologique précis permettant d'évaluer rigoureusement certaines thèses centrales du déflationnisme aléthique contemporain.

Dans une première partie, nous nous sommes efforcés de donner un sens clair à l'idée, souvent attribuée aux déflationnistes, selon laquelle la vérité serait une sorte de notion logique. En nous appuyant sur les analyses et les outils issus de la tradition inférentialiste, nous avons pu tracer une frontière délimitant le champ des notions proprement logiques. Dans ce cadre, nous avons montré que le prédicat de vérité, même caractérisé par des règles minimales, ne pouvait guère être qualifié de logique. Le principal problème concerne ici la nature des noms (canoniques) des énoncés auxquels le prédicat de vérité est censé s'appliquer : dans le passage de l'énoncé « «$_c$ la neige est blanche $_c$» [306] est vrai » à l'énoncé « la neige est blanche » (et inversement), qui peut se traduire par une règle d'élimination (ou par une règle d'introduction pour l'inférence inverse), il est fait un usage crucial d'une information de nature sémantique, à savoir la référence du nom canonique présent dans l'énoncé contenant le prédicat de vérité. Bien plus, le contenu sémantique de l'énoncé « la neige est blanche » semble encore actif dans l'énoncé « «$_c$ la neige est blanche $_c$» est vrai », alors même qu'il n'y est plus directement présent. Les noms canoniques paraissent donc posséder une nature hybride : si on peut être tenté de les ranger dans la catégorie grammaticale des noms propres, ils doivent néanmoins posséder une certaine « transparence » qui permette de retrouver, à la simple donnée du nom lui-même, l'énoncé auquel il réfère. Cette dernière caractéristique les rapproche de phénomènes linguistiques complexes connus sous le nom d'autonymie. Nous avons vu que par conséquent l'emploi de ces noms canoniques dans les règles d'introduction et d'élimination pour « vrai » contredisait la pure structuralité de ces règles et les conduisait également à briser la contrainte d'harmonie globale. Or, ces critères de structuralité et d'harmonie forment deux caractéristiques essentielles et incontournables de la logicité selon l'approche inférentialiste. Nous avons donc conclu que la vérité n'était pas une notion logique à la lumière des critères inférentialistes de logicité. Malgré ces résultats négatifs, la question de la logicité de la vérité demeure en partie ouverte et pourra faire l'objet de recherches supplémentaires. Nous avons en effet employé ici un cadre méthodologique et une interprétation de la logicité issus de la tradition inférentialiste. Néanmoins, l'établissement d'une démarcation délimitant le domaine des notions logiques est notoirement problématique, et, en parallèle de la tradition inférentialiste, d'autres démarches sont possibles [307]. En s'appuyant sur d'autres critères de logicité, on pourra poursuivre l'exa-

306. Pour rappel : nous avons noté «$_c$ p $_c$» le nom canonique d'un énoncé p sans préjuger de la nature de ce nom ni de la manière dont il s'obtient.
307. Pour un panorama général des diverses approches possibles, voyez MacFarlane (2017).

men précis de la nature plus ou moins logique de la vérité déflationniste. Ces approches sont en plein développement et nous semblent être complémentaires de celle proposée ici [308].

La seconde partie de notre travail fut consacrée à l'analyse d'un argument anti-déflationniste —dû, de manière indépendante, à Ketland (1999) et à Shapiro (1998b)— qui a suscité d'importantes discussions au cours des vingt dernières années. Cette fois, l'attention se porte principalement sur l'absence de contenu (explicatif) propre de la notion de vérité, et le cadre méthodologique est celui des extensions aléthiques d'une théorie de base capable d'exprimer sa propre syntaxe [309]. Selon l'argument, les allégations de « non-substantialité » ou d'absence de contenu que les déflationnistes prononcent à l'endroit de la vérité devraient se traduire par une propriété de conservativité. Mais cette contrainte de conservativité de notre théorie de la vérité sur une théorie de base semble incompatible avec la formalisation de certains raisonnements sémantiques développés à la suite des théorèmes d'incomplétude de Gödel et au cœur desquels on trouve une dérivation d'un principe de réflexion appuyée sur la notion de vérité. Les débats qui ont suivi la publication de cet argument s'articulent autour de trois axes que nous avons examinés tour à tour et qui correspondent à autant de stratégies de réponse possibles pour le déflationniste.

Dans un premier temps, nous sommes revenus sur la contrainte de conservativité et avons défendu l'idée qu'une telle contrainte devait bien s'appliquer aux théories déflationnistes de la vérité. Nous n'avons bien sûr pu que constater l'absence de consensus sur cette question, y compris dans les rangs des déflationnistes. Mais, selon nous, la comparaison avec d'autres exemples historiques d'emplois de la conservativité dans des tentatives visant à « dégonfler » des concepts [310] suggère que la conservativité est une manière naturelle de traduire dans un cadre formel précis les thèses déflationnistes concernant l'absence de contenu et de rôle explicatif propres à la vérité. Si la vérité doit bénéficier d'un régime d'exception en la matière, c'est sans doute aux déflationnistes

308. Pour un travail récent sur cette question et qui propose une défense de la logicité de la vérité appuyée sur une caractérisation de la logique au moyen de propriétés d'invariance, voyez Bonnay et Galinon (2018).

309. Pour rappel : nous appelons extension aléthique d'une théorie de base T exprimée dans un langage $\mathcal{L}_T$, une théorie T' étendant T ($T \subset T'$) qui est exprimée dans un langage $\mathcal{L}_{T'}$ étendant $\mathcal{L}_T$ au moyen d'un nouveau prédicat Vr ($\mathcal{L}_T \subset \mathcal{L}_T \cup \{Vr\} \subseteq \mathcal{L}_{T'}$). T' se compose de (nouveaux) axiomes censés gouverner le prédicat de vérité Vr pour $\mathcal{L}_T$.

310. Nous pensons ici principalement aux exemples du Programme de Hilbert et du fictionnalisme mathématique de Field.

que revient la double tâche d'expliquer plus précisément pourquoi et de fournir un autre critère permettant de tracer la frontière entre les notions purement expressives et celles possédant un réel rôle explicatif. Dans cette partie, nous nous sommes également attachés à préciser sur quel type de théories de base une théorie déflationniste de la vérité doit être conservative : plutôt que d'exiger que la vérité soit conservative sur toute théorie ou sur la logique pure, il semble plus raisonnable de privilégier une contrainte de conservativité sur une théorie de base déjà munie de sa propre syntaxe. Quoi qu'il en soit, si la contrainte de conservativité demeure controversée, elle s'est néanmoins imposée de fait comme un des thèmes dominants des recherches actuelles sur le déflationnisme qui entendent s'appuyer sur les outils précis de la logique et de la philosophie formelle. L'étude des propriétés de conservativité de diverses théories axiomatiques de la vérité est aujourd'hui un thème de recherche très actif. Par rapport aux résultats que nous avons retenus dans notre travail, les principaux développement en cours concernent les propriété de conservativité de théories de la vérité non typée, *i.e.* les systèmes aléthiques dans lesquels le prédicat de vérité peut s'appliquer à des énoncés qui contiennent déjà le prédicat de vérité, ainsi que l'étude des propriétés de conservativité sur des théories arithmétiques plus faibles que l'arithmétique de Peano [311].

Le second axe de discussion que nous avons examiné a été ouvert par la réponse de Field [312] à l'argument de Ketland et Shapiro. Field tente de montrer que la non-conservativité d'une extension aléthique tarskienne sur l'arithmétique de Peano au premier ordre est compatible avec les canons déflationnistes. Pour cela, il distingue les axiomes « essentiels » de la vérité, censés être conservatifs, et attribue la non-conservativité uniquement à l'extension des axiomes d'induction à un langage plus riche que celui de l'arithmétique. Malgré l'ingéniosité de sa démarche, nous montrons cependant que la ligne d'argumentation de Field n'est pas réellement tenable. Un examen approfondi de la démonstration sémantique établissant la cohérence de PA au sein d'une extension aléthique tarskienne montre que les axiomes sémantiques ont déjà un impact avant même l'extension de l'induction et qu'ils jouent un rôle essentiel et indispensable dans cette preuve. Premièrement, l'extension de l'induction au langage contenant le prédicat de vérité ne suffit pas à obtenir une extension non-conservative ; les clauses sémantique

311. Sur ces points, les travaux de Cieśliński (2017), Halbach (2014), Horsten (2011) et Horsten et Leigh (2017) ont été et demeurent capitaux. Cieśliński (2017) expose d'ailleurs plusieurs problèmes ouverts sur ces questions.
312. Field (1999)

tarskiennes sont donc bel et bien indispensables pour obtenir une preuve de $Con(PA)$. Deuxièmement, nous avons rappelé que le passage de PA à PA_{Tar} [313] a déjà un impact sur la structure des modèles possibles puisque, sans induction élargie, PA_{Tar} n'est pas modèle-théoriquement conservative sur PA. Nous en avons ensuite conclu que la conservativité *syntaxique* de PA_{Tar} sur PA n'est pas due à la modestie de la vérité mais bien au fait qu'en l'occurence pouvoir explicatif et pouvoir expressif sont indissociables. Enfin, une comparaison fine de la preuve sémantique de cohérence avec une preuve de cohérence obtenue dans un sous-fragment de l'arithmétique du second ordre [314] montre que les axiomes compositionnels sont centraux pour traduire la relation ensembliste d'appartenance et permettent d'exprimer l'équivalent d'axiomes de compréhension forts. Dès lors, malgré les tentatives fieldiennes, il nous paraît impossible d'exonérer la vérité de toute « responsabilité » ou de tout contenu explicatif dans la preuve sémantique tarskienne de cohérence. Cette stratégie de réponse nous semble donc déboucher sur une impasse.

Le troisième et dernier axe de discussion est celui qui nous semble offrir les perspectives les plus prometteuses. Il consiste à prendre au sérieux la contrainte de conservativité et à renoncer aux preuves sémantiques de cohérence. Ce faisant, le déflationniste doit montrer comment il peut se passer de ce type de raisonnements en indiquant comment il peut les reconstruire ou les remplacer sans s'appuyer sur une notion forte de vérité. Plus précisément, le problème revient alors essentiellement à fournir une justification (non sémantique) d'un schéma de réflexion pour une théorie de base donnée. L'analyse se concentre alors sur ce en quoi consiste pour un agent soumis à des contraintes de rationalité qu'accepter une théorie et sur la mise au jour des éventuels engagements implicites que cet agent contracte lorsqu'il accepte une théorie. Si, nous l'avons dit, cette voie de recherche nous semble prometteuse, les résultats contenus dans notre travail incitent néanmoins à la prudence. En premier lieu, parler d'un « processus de réflexion » sur notre acceptation d'une théorie sans plus de précisions pour justifier tel ou tel schéma de réflexion apparaît insuffisant. Nous avons exposé des modèles plausibles d'acceptation qui ne se conforment pas à cette analyse, notamment un pour lequel l'adoption d'un

313. Pour rappel : PA désigne ici l'arithmétique de Peano formulée dans le langage $\mathcal{L}_{PA}$ tandis que PA_{Tar} désigne la théorie formulée dans $\mathcal{L}_{PA} \cup Vr$ obtenue en adjoignant à PA des clauses récursives tarskiennes pour un prédicat « Vr » mais sans étendre l'induction aux énoncés contenant ce prédicat nouvellement introduit.

314. Cette comparaison est possible car PA^{+Ind}_{Tar} et le sous-fragment ACA sont preuve-théoriquement équivalentes. Pour plus de détails sur ce résultat classique *cf.* Halbach (2014).

schéma de réflexion semble aboutir à une contradiction méthodologique. Par ailleurs, les tentatives de dérivation ou de justification détaillée d'un schéma de réflexion pour une théorie de base (ou même simplement d'un énoncé exprimant sa cohérence) à partir d'une notion forte d'acceptation, telle que l'acceptation rationnelle ou l'acceptation justifiée, ne sont pas sans risques. Pour une d'entre elles[315], nous avons pu montrer qu'elle se soldait par un échec dans la mesure où elle entendait s'appuyer sur un principe de rationalité fort, baptisé Principe de Responsabilité en première personne, qui s'il pouvait paraître plausible à première vue s'est finalement révélé intenable. Une autre tentative plus récente[316] semble, quant à elle, couronnée d'un plus franc succès. Elle repose néanmoins sur une notion de crédibilité dont l'interprétation n'est pas encore totalement éclaircie. La nature du processus de réflexion permettant de justifier l'adoption de schémas de réflexion pour l'arithmétique ou pour d'autres théories reste donc à ce jour mal connue. L'examiner plus en profondeur et tenter de la clarifier constituent l'une des tâches auxquelles devront se confronter dans les années à venir les philosophes intéressés par le déflationnisme en matière vérité.

Au final, si —sans doute est-il temps de le confesser— nous ne sommes pas nous-même partisans du déflationnisme aléthique, nous devons reconnaître qu'au sortir de notre exploration, ce dernier apparaît loin d'être désarmé. L'une des difficultés qui ont accompagné l'avénement du déflationnisme contemporain sur le devant de la scène philosophique aura été de traduire précisément les thèses déflationnistes, souvent exprimées de façon informelle, de manière à pouvoir en évaluer le plus exactement possible la solidité et la cohérence. Dans notre travail, nous avons distingué deux paradigmes possibles pour tenter de surmonter cet obstacle. Le premier a consisté à étudier la logicité du prédicat de vérité, le second s'est attaché à examiner les propriétés de conservativité de théories axiomatisant la vérité. Après l'avoir soumis à un examen serré dans ce double cadre méthodologique, nous avons pu jeter une lumière plus claire sur la nature du déflationnimse aléthique. Sans doute certaines thèses déflationnistes méritent être nuancées : une interprétation trop stricte de la logicité de la vérité nous paraît irrecevable, certaines voies de réponse initiées à la suite des arguments de Ketland et Shapiro nous semblent devoir être abandonnées. Toutefois, rien de ce que nous avons dit ne vaut réfutation du déflationnisme aléthique. Et, si l'on est prêt à embrasser toutes les conséquences méthodologiques de cette position —en particulier, si l'on est prêt à renoncer aux preuves

315. Celle proposée dans Galinon (2014, 2010).
316. Il s'agit de celle développée dans Cieśliński (2017).

sémantiques de cohérence— le déflationnisme apparaît comme une position originale et viable en matière de philosophie du langage et de la logique. Aujourd'hui encore, c'est un champ de recherche très actif et son succès ne s'est pas démenti depuis que nous avons commencé notre travail.

Bibliographie

[1] William P. Alston. « The Deontological Conception of Epistemic Justification ». In : *Philosophical Perspectives* 2, Epistemology (1988), p. 257–299 (cf. p. 388).

[2] Denis Apothéloz. *Rôle et fonctionnement de l'anaphore dans la dynamique textuelle*. Langues et cultures 29. Genève : Librairie Droz, 1995 (cf. p. 27).

[3] Aristote. *Metaphysique*. Traduction de Jules Tricot. Vrin (cf. p. 41).

[4] Bradley Armour-Garb. « Challenges to Deflationary Theories of Truth ». In : *Philosophy Compass* 7.4 (2012), p. 256–266. issn : 1747-9991. doi : 10.1111/j.1747-9991.2011.00462.x. url : http://dx.doi.org/10.1111/j.1747-9991.2011.00462.x (cf. p. 206).

[5] Bradley Armour-Garb et J.C. Beall, éds. *Deflationary Truth*. Open Court, 2005 (cf. p. 443).

[6] Alfred Jules Ayer. *Language, Truth and Logic*. 2e éd. Traduction française (Ohana, 1956). New York : Dover Publications, 1946 (cf. p. 23–24, 450).

[7] Alfred Jules Ayer. « The Criterion of Truth ». In : *Analysis* 3 (1935) (cf. p. 96).

[8] Jody Azzouni. « Comments on Shapiro ». In : *The Journal of Philosophy* 96.10 (1999), p. 541–544 (cf. p. 294, 339).

[9] Robert Batterman. « Intertheory Relations in Physics ». In : *The Stanford Encyclopedia of Philosophy (Fall 2012 Edition), Edward N. Zalta (ed.)* (2012). url : http://plato.stanford.edu/archives/fall2012/entries/physics-interrelate/ (cf. p. 379).

[10] Timothy Bays. « Beth's Theorem and Deflationism ». In : *Mind* 118.472 (2009), p. 1061–1073 (cf. p. 170, 301).

[11] J.C. Beall et Bradley Armour-Garb, éds. *Deflationism and Paradox*. Oxford University Press, 2005 (cf. p. 5).

[12] Lev D. Beklemishev. « Reflection principles and provability algebras in formal arithmetic ». In : *Russian Mathematical Survey* 60.2 (2005), p. 197–268 (cf. p. 341–343, 370–371).

[13] Nuel D. Belnap. « Tonk, plonk and plink ». In : *Analysis* 22 (1962), p. 130–134 (cf. p. 141).

[14] Simon Blackburn et Keith Simmons, éds. *Truth*. Oxford Readings in Philosophy. Oxford University Press, 1999 (cf. p. 95, 443).

[15] Robert Blanché et Jacques Dubucs. *La Logique et son histoire*. Armand Colin, 1996 (cf. p. 210).

[16] Denis Bonnay et Michaël Cozic, éds. *Philosophie de la logique : conséquence, preuve et vérité*. Vrin, 2009 (cf. p. 439).

[17] Denis Bonnay et Henri Galinon. « Deflationary Truth is a Logical Notion ». In : *Truth, Existence and Explanation : FilMat 2016 studies in the philosophy of mathematics*. Sous la dir. de Mario Piazza et Gabriele Pulcini. T. 334. Boston Studies in the Philosophy and History of Science. Springer Verlag, 2018, p. 71–88 (cf. p. 427).

[18] George S. Boolos. *The Logic of Provability*. Cambridge University Press, 1993 (cf. p. 370–371, 373).

[19] George S. Boolos, John P. Burgess et Richard C. Jeffrey. *Computability and Logic*. 4e éd. Cambridge University Press, 2002 (cf. p. 255).

[20] Philippe de Brabanter. « Philosophie du langage et autonymie : une déjà longue histoire ». In : *Histoire, Épistémologie, Langage* 27 (2005), p. 1–30. URL : www.institutnicod.org (cf. p. 177, 179, 181, 184, 194–195).

[21] Robert Brandom. *Making it Explicit*. Cambridge, Massachusetts : Harvard University Press, 1994 (cf. p. 96).

[22] Robert Brandom. « Pragmatism, Phenomenalism and Truth Talk ». In : *Midwest Studies in Philosophy*. Sous la dir. de P. French, T. Uehling et H. Wettstein. T. 12. Minneapolis, Minnesota : University of Minnesota Press, 1988 (cf. p. 96).

[23] Franz Brentano. *Wahrheit und Evidenz*. Leipzig : Felix Meiner, 1930 (cf. p. 31).

[24] Tyler Burge. *Cognition Through Understanding : Self-knowledge, Interlocution, Reasoning Reflection : Philosophical Essays*. T. 3. Oxford University Press, 2013 (cf. p. 403).

[25] Tyler Burge. « Computer Proof, Apriori Knowledge, and Other Minds ». In : *Noûs* 32.Supplement 12 (1998), p. 1–37. doi : 10.1111/0029-4624.32.s12.1 (cf. p. 403).

[26] Tyler Burge. « Content preservation ». In : *The Philosophical Review* 102.4 (1993), p. 457–488 (cf. p. 403).

[27] Tyler Burge. « Our entitlement to self-knowledge : I. Tyler Burge ». In : *Proceedings of the Aristotelian Society, New Series* 96 (1996), p. 91–116 (cf. p. 392).

[28] Tyler Burge. « Perceptual entitlement ». In : *Philosophy and Phenomenological Research* 67.3 (2003), p. 503–548 (cf. p. 403).

[29] John P. Burgess. « On the outside looking in : A caution about conservativeness ». In : *Kurt Gödel : Essays for His Centennial*. Sous la dir. de Solomon Feferman, Charles Parsons et Stephen George Simpson. Lecture Notes in Logic. New York : Cambridge University Press, 2010, p. 128–141 (cf. p. 278).

[30] John P. Burgess et Gideon Rosen. *A Subject With No Object: Strategies for Nominalistic Interpretations of Mathematics*. Oxford, Clarendon Press, 1997 (cf. p. 253, 277, 294).

[31] Stewart Candlish et Nic Damnjanovic. « A Brief History of Truth ». In : *Philosophy of Logic*. Sous la dir. de Dale Jacquette. Handbook of the Philosophy of Science. North-Holland Publishing Company, 2007, p. 227–323 (cf. p. 1).

[32] Herman Cappelen et Ernest Lepore. « Quotation ». In : *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition), Edward N. Zalta (ed.)* (2012). url : http://plato.stanford.edu/archives/spr2012/entries/quotation/ (cf. p. 177, 179, 182, 184, 195).

[33] Cezary Cieśliński. *The Epistemic Lightness of Truth : Deflationism and its Logic*. Cambridge University Press, 2017 (cf. p. 270–276, 366, 380, 403–413, 415–419, 428, 430, 444).

[34] Cezary Cieśliński. « Truth, conservativeness and provability ». In : *Mind* 119.474 (2010), p. 409–422 (cf. p. 362–366, 368–371, 377, 380–381, 385–386).

[35] William Kingdon Clifford. « The ethics of belief ». In : *Contemporary Review* (1877) (cf. p. 388).

[36] L. Jonathan Cohen. « Belief and acceptance ». In : *Mind* 98.391 (1989), p. 367–389 (cf. p. 387).

[37] Mark Colyvan. *The Indispensability of Mathematics*. Oxford University Press, 2001 (cf. p. 220–222).

[38] John Corcoran, William Frank et Michael Maloney. « String Theory ». In : *Journal of Symbolic Logic* 39.4 (déc. 1974), p. 625–637 (cf. p. 240, 291).

[39] Mikaël Cozic. « Introduction et présentation de Tarski (1944) ». In : *Philosophie de la logique : conséquence, preuve et vérité*. Sous la dir. de Denis Bonnay et Mikaël Cozic. Textes clés. Vrin, 2009, p. 247–253 (cf. p. 68).

[40] William Craig et Robert Vaught. « Finite axiomatizability using additional predicates ». In : *Journal of Symbolic Logic* 23 (1958), p. 289–308 (cf. p. 309).

[41] Haskell B. Curry. « A formalization of recursive arithmetic ». In : *American Journal of Mathematics* 63.2 (1941), p. 263–282 (cf. p. 213).

[42] Nic Damnjanovic. « Deflationism and the Success Argument ». In : *The Philosophical Quarterly* 55.218 (2005), p. 53–67 (cf. p. 285–287).

[43] Nic Damnjanovic. « New Wave Deflationism ». In : *New Waves in Truth*. Sous la dir. de Cory D. Wright et Nikolaj J. L. L. Pedersen. Pgrave-Macmillan, 2010. Chap. 3, p. 45–58 (cf. p. 7).

[44] Donald Davidson. *Enquêtes sur la vérité et l'interprétation (traduit de l'anglais par Pascal Engel)*. Rayon Philo. Nîmes : Éditions Jacqueline Chambon, 1993 (cf. p. 436).

[45] Donald Davidson. *Inquiries into Truth and Interpretation : Philosophical Essays*. T. 2. Traduction française par Pascal Engel *in* (Davidson, 1993). Oxford, UK : Oxford University Press, 1984 (cf. p. 50).

[46] Donald Davidson. « Quotation ». In : *Theory and Decision* 11 (1979), p. 27–40 (cf. p. 179, 194).

[47] Walter Dean. « Arithmetical Reflection and the Provability of Soundness ». In : *Philosophia Mathematica* 23.1 (2015), p. 31–64. doi : 10.1093/philmat/nku026 (cf. p. 361, 376).

[48] Michael Detlefsen. *Hilbert's program: an essay on mathematical instrumentalism*. T. 182. Synthese Library. Springer, 1986 (cf. p. 219).

[49] Michael Detlefsen. « On an alleged refutation of Hilbert's Program using Gödel's first incompleteness theorem ». In : *Journal of Philosophical Logic* 19 (1990), p. 343–377 (cf. p. 211, 219).

[50] Michael Detlefsen. « On interpreting Gödel's second theorem ». In : *Journal of Philosophical Logic* 8 (1979), p. 297–313 (cf. p. 219).

[51] Michael Detlefsen. « What Does Gödel's Second Theorem Say ? » In : *Philosophia Mathematica* 9.3 (2001), p. 37–71 (cf. p. 219).

[52] David DeVidi et Graham Solomon. « Tarski on "essentially richer" metalanguages ». In : *Journal of Philosophical Logic* 28 (1999), p. 1–28 (cf. p. 54, 62–63, 251).

[53] Michael Devitt. « The metaphysics of deflationary truth ». In : *What is Truth ?* Sous la dir. de Richard Schantz. De Gruyter, 2001, p. 60–78 (cf. p. 79).

[54] *Dictionnaire de français Larousse*. 2018. URL : https://www.larousse.fr/dictionnaires/francais/alÃĪthique/2184 (cf. p. 2).

[55] Michael Dummett. *Frege : Philosophy of Language*. Londres : Duckworth, 1973 (cf. p. 148).

[56] Michael Dummett. *The Logical Basis of Metaphysics*. Cambridge, Massachusetts : Harvard University Press, 1991 (cf. p. 141, 148, 151).

[57] Michael Dummett. « The Philosophical Significance of Gödel's Theorem ». In : *Ratio* 5 (1963). repris in Dummett (1978), p. 140–155 (cf. p. 349).

[58] Michael Dummett. *Truth and Other Enigmas*. Duckworth, 1978 (cf. p. 238, 437).

[59] Michael J. Dunn et Nuel D. Belnap. « The Substitution Interpretation of the Quantifiers ». In : *Noûs* 2.2 (1968), p. 177–185 (cf. p. 34).

[60] Ali Enayat et Albert Visser. « New Constructions of Satisfaction Classes ». In : *Unifying the Philosophy of Truth*. Sous la dir. de T. Achourioti et al. T. 36. Logic, epistemology and the unity of science. Dordrecht, Springer, 2015, p. 321–335 (cf. p. 307, 317).

[61] *Encyclopédie Universalis*. 2018. URL : https://www.universalis.fr/dictionnaire/alethique/ (cf. p. 2).

[62] Pascal Engel, éd. *Believing and Accepting*. Kluwer Academic Publishers, 2000 (cf. p. 387).

[63] Fredrik Engström. « Satisfaction Classes in Non-Standard Models of First-Order Arithmetic ». Mém.de mast. Chalmers University of Technology et Göteborg University, 2002 (cf. p. 316).

[64] William Ewald, éd. *From Kant to Hilbert : A Source book in the Foundations of Mathematics*. T. II. Oxford, Clarendon Press, 1996 (cf. p. 444).

[65] Solomon Feferman. « Arithmetization of metamathematics in a general setting ». In : *Fundamenta Mathematicae* 49 (1960), p. 35–92 (cf. p. 341).

[66] Solomon Feferman. *In The Light of Logic*. Oxford University Press, 1998 (cf. p. 438).

[67] Solomon Feferman. « Kurt Gödel : Conviciton and Caution ». In : *In The Light of Logic (*Feferman*, 1998a)*. Oxford University Press, 1998, p. 150–164 (cf. p. 249).

[68] Solomon Feferman. « Reflecting on Incompleteness ». In : *Journal of Symbolic Logic* 56.1 (1991), p. 1–49 (cf. p. 77, 329, 353, 387).

[69] Solomon Feferman. « Transfinite Recursive Progressions of Axiomatic Theories ». In : *Journal of Symbolic Logic* 27 (1962), p. 259–316 (cf. p. 341–342, 387).

[70] Solomon Feferman et al., éds. *Kurt Gödel - Collected Works*. T. I (Publications 1929-1936). Oxford University Press, 1986 (cf. p. 442).

[71] Luis Fernández-Moreno. « Tarskian Truth And The Correspondence Theory ». In : *Synthese* 126.1 (2001), p. 123–148 (cf. p. 78).

[72] Hartry Field. « Attributions of Meaning and Content ». In : *Truth and the Absence of Fact*. Oxford University Press, 2001. Chap. 5, p. 157–174 (cf. p. 129).

[73] Hartry Field. « Critical Notice : Paul Horwich's *Truth* ». In : *Philosophy of Science* 59 (1992), p. 321–330 (cf. p. 103, 109, 135).

[74] Hartry Field. « Deflating the Conservativeness Argument ». In : *The Journal of Philosophy* 96.10 (1999), p. 533–540 (cf. p. 206, 265, 276, 281, 294–298, 301, 309, 318–319, 323, 331, 335, 345, 365, 428).

[75] Hartry Field. « Deflationist Views of Meaning and Content ». In : *Mind* 103 (1994). Repris avec un Postscript *in* Field, 2001b, p. 104-156, p. 249–85 (cf. p. 9, 96, 114–122, 124–128, 192, 235, 253).

[76] Hartry Field. « Disquotational Truth and Factually Defective Discourse ». In : *The Philosophical Review* 103.3 (1994). Repris *in* Field, 2001b, p. 222-258 (cf. p. 114, 122–123, 126, 134, 192, 235, 253).

[77] Hartry Field. « Mathematics without truth (a reply to Maddy) ». In : *Pacific Philosophical Quarterly* 71 (1990), p. 206–222 (cf. p. 224).

[78] Hartry Field. « Mental Representation ». In : *Erkenntnis* 13.1 (1978). Repris avec un Postscript *in* Field, 2001b, p. 30-82 (cf. p. 109, 111–112).

[79] Hartry Field. « On conservativeness and incompleteness ». In : *Journal of Philosophy* 82.5 (1985), p. 239–260 (cf. p. 222, 224).

[80] Hartry Field. « Quine and the Correspondance Theory ». In : *The Philosophical Review* 83.2 (1974). Repris avec un Postscript *in* Field, 2001b, p. 199-221 (cf. p. 109, 112).

[81] Hartry Field. « Reply to Anil Gupta and José Martínez-Fernández ». In : *Philosophical Studies* 124.1 — Symposium sur Truth and the Absence of Fact (2005), p. 105–110 (cf. p. 129).

[82] Hartry Field. « Reply to Barry Loewer ». In : *Philosophical Studies* 124.1 — Symposium sur Truth and the Absence of Fact (2005), p. 110–118 (cf. p. 129).

[83] Hartry Field. « Reply to Vann McGee ». In : *Philosophical Studies* 124.1 — Symposium sur Truth and the Absence of Fact (2005), p. 118–128 (cf. p. 129).

[84] Hartry Field. *Science Without Numbers*. Library of Philosophy and Logic. Oxford Basil Blackwell, 1980 (cf. p. 219–221, 223, 225, 240, 279).

[85] Hartry Field. « Tarski's theory of truth ». In : *The Journal of Philosophy* 69.13 (1972). Repris avec un Postscript *in* Field, 2001b, p. 3-29, traduction française et présentation *in* Bonnay et Cozic, 2009, p. 347–375 (cf. p. 109–113, 173).

[86] Hartry Field. « The Deflationary Conception of Truth ». In : *Fact, Science and Morality*. Sous la dir. de Graham MacDonald et Crispin Wright. Oxford : Basil Blackwell, 1986, p. 55–117 (cf. p. 14, 16, 96, 285).

[87] Hartry Field. « Theory change and the indeterminacy of reference ». In : *The Journal of Philosophy* 70.14 (1973). Repris avec un Postscript *in* Field, 2001b, p. 177-198 (cf. p. 109, 112).

[88] Hartry Field. « Truth and the Absence of Fact — Precis ». In : *Philosophical Studies* 124.1 — Symposium sur Truth and the Absence of Fact (2005), p. 41–44 (cf. p. 129).

[89] Hartry Field. *Truth and the Absence of Fact*. Oxford University Press, 2001 (cf. p. 109, 112–114, 122, 125, 129, 134–135, 439–440).

[90] Arthur Fine. « The Natural Ontological Attitude ». In : *Scientific Realism*. Sous la dir. de J. Leplin. Berkeley, California : University of California Press, 1984 (cf. p. 96).

[91] Torkel Franzén. « Transfinite Progression : a second look at completeness ». In : *The Bulletin of Symbolic Logic* 10.3 (2004), p. 367–389 (cf. p. 340, 405–407, 417).

[92] Gottlob Frege. *Écrits logiques et philosophiques*. Traduction et introduction de Claude Imbert. Paris, France : Seuil, 1971 (cf. p. 440).

[93] Gottlob Frege. *Écrits posthumes*. Sous la dir. de Philippe de Rouilhan et Claudine Tiercelin. Traduction française de Frege (1969). Nîmes : Jacqueline Chambon, 1999 (cf. p. 14, 440).

[94] Gottlob Frege. *Funktion und Begriff*. Texte d'une conférence prononcée devant la *Société savante d'Iéna pour la médecine et les sciences naturelles*. Traduction française « Fonction et concept » *in* (Frege, 1971). 1891 (cf. p. 96).

[95] Gottlob Frege. « Logische Untersuchungen ». In : *Beiträge zur Philosophie des deutsches Idealismus* 1, 3 ((1918-1923)). Traduction française « Recherches logiques » *in* (Frege, 1971) (cf. p. 10, 12, 96).

[96] Gottlob Frege. *Mes intuitions logiques fondamentales*. Posthume. Traduction française *in* (Frege, 1999). 1915 (cf. p. 13–14).

[97] Gottlob Frege. *Nachgelassene Schriften*. Sous la dir. d'Hans Hermes, Friedrich Kambartel et Friedrich Kaulbach. Hambourg : Felix Meiner, 1969 (cf. p. 440).

[98] Gottlob Frege. « Über Sinn und Bedeutung ». In : *Zeitschrift für Philosophie und philosophische Kritk* 100 (1892). Traduction française « Sens et dénotation » *in* (Frege, 1971) (cf. p. 8–9).

[99] Greg Frost-Arnold. « Tarski's nominalism ». In : *New Essays on Tarski and Philosophy*. Sous la dir. de Douglas Patterson. Oxford University Press, 2008. Chap. 9, p. 225–246 (cf. p. 41).

[100] Henri Galinon. « Acceptation, cohérence et responsabilité ». In : *Liber Amicorum Pascal Engel*. Sous la dir. de Julien Dutant, Davide Fassio et Anne Meylan. Université de Genève, 2014, p. 320–333. url : http://www.unige.ch/lettres/philo/publications/engel/liberamicorum (cf. p. 384–388, 390–393, 395–396, 398, 400, 402, 430).

[101] Henri Galinon. « Deflationary truth : conservativity or logicality ? » In : *The Philosophical Quarterly* 65.259 (2015), p. 268–274. doi : 10.1093/pq/pqu087 (cf. p. 280).

[102] Henri Galinon. « Déflationnisme et conservativité : quelqu'un a-t-il changé de sujet ? » In : *Philosophia Scientiae* 16.3 (2012), p. 133–151 (cf. p. 280, 287).

[103] Henri Galinon. « Recherches sur la Vérité—Définition, Élimination, Déflation ». Thèse de doct. Université Paris 1 Panthéon-Sorbonne, 2010 (cf. p. 9, 113, 147–149, 155, 158, 160–165, 175, 184, 266, 280–281, 287, 380, 384–388, 390–396, 398–400, 402, 430).

[104] Peter Geach. « Assertion ». In : *The Philosophical Review* 74.4 (1965), p. 449–465 (cf. p. 11).

[105] Peter Geach. « Intentional Identity ». In : *Journal of Philosophy* 64.20 (1967), p. 627–632 (cf. p. 29).

[106] Peter Geach. *Mental Acts. Their Content and Their Objects*. Londres : Routledge & Kegan Paul Ltd, 1957 (cf. p. 181).

[107] Gerhard Gentzen. *Recherches sur la déduction logique (traduction et commentaires de* Gentzen*, 1935, Robert Feys & Jean Ladrière)*. Paris, France : Presses Universitaires de France, 1955 (cf. p. 138–139, 441).

[108] Gerhard Gentzen. « Untersuchungen über das logische Schliessen ». In : *Mathematische Zeitschrift* 39.1 (1935). Traduction française *in* (Gentzen, 1955)., p. 176–210 (cf. p. 138, 441).

[109] Steven Givant. « Bibliography of Alfred Tarski ». In : *Journal of Symbolic Logic* 51.4 (déc. 1986), p. 913–941 (cf. p. 37).

[110] Paul Gochet. *Ascent to Truth - A Critical Examination of Quine's Philosophy*. München Wien : Philosophia Verlag, 1986 (cf. p. 84, 90).

[111] Paul Gochet. *Quine en perspective*. Nouvelle bibliothèque scientifique. Flammarion, 1978 (cf. p. 85).

[112] Kurt Gödel. « Über formal unentscheidbare Sätze de *Principia mathematica* und verwandter Systeme I ». In : *Monatshefte für Mathematik und Physik* 38 (1931). Traduction anglaise *in* Feferman et al., 1986, traduction française in ; Nagel, Newman et Girard, 1989, p. 173–198 (cf. p. 244, 249).

[113] Nelson Goodman et Willard Van Orman Quine. « Steps toward a constructive nominalism ». In : *The Journal of Symbolic Logic* 12.4 (1947), p. 105–122 (cf. p. 220).

[114] Reuben Goodstein. « Logic-free formalisations of recursive arithmetic ». In : *Mathematica Scandinavica* 2 (1954), p. 247–261 (cf. p. 213).

[115] Mitchell Green et John Williams, éds. *Moore's Paradox : New Essays on Belief, Rationality and the First-Person*. New York : Oxford University Press, 2007 (cf. p. 393).

[116] Mark Greenberg et Glibert Harman. « Conceptual Role Semantics ». In : *The Oxford Handbook of Philosophy of Language*. Sous la dir. d'Ernest Lepore et Barry Smith. Oxford, UK : Oxford University Press, 2008, p. 296–322 (cf. p. 117).

[117] Dorothy Grover. *A Prosentential Theory of Truth*. Princeton, New Jersey : Princeton University Press, 1992 (cf. p. 25).

[118] Dorothy Grover. « Propositional Quantification and Quotation Contexts ». In : *Truth Syntax and Modality*. Sous la dir. d'Hugues Leblanc. T. 68. Studies in Logic and the Foundations of Mathematics. North-Hollandh, 1973 (cf. p. 25).

[119] Dorothy Grover. « Propositional Quantifier ». In : *Journal of Phiosophical Logic* 1.2 (1972), p. 111–136 (cf. p. 25).

[120] Dorothy Grover, Joseph Camp et Nuel D. Belnap. « A Prosentatial Theory of Truth ». In : *Philosophical Studies* 27.2 (1975), p. 73–125 (cf. p. 25–27, 29–31, 33–36, 96, 101, 113).

[121] Anil Gupta. « A Critique of Deflationism ». In : *Philosophical Topics* 21 (1993). Repris *in* Armour-Garb et Beall, 2005 et *in* ; Blackburn et Simmons, 1999, p. 57–81 (cf. p. 74, 235, 296, 403).

[122] Anil Gupta et José Martínez-Fernández. « Field on the Concept of Truth — Comment ». In : *Philosophical Studies* 124.1 — Symposium sur Truth and the Absence of Fact (2005), p. 45–58 (cf. p. 120, 122).

[123] Volker Halbach. *Axiomatic Theories of Truth*. 2e éd. Cambridge, UK : Cambridge University Press, 2014. doi : 10.1017/CBO9781139696586 (cf. p. 77, 153, 201, 257, 270, 304–305, 316–317, 319, 322–323, 329–330, 334–335, 341–342, 344–345, 353, 356, 428–429).

[124] Volker Halbach. « Conservative Theories of Classical Truth ». In : *Studia Logica* 62.3 (1999), p. 353–370 (cf. p. 254, 257, 317, 322, 334).

[125] Volker Halbach. « Disquotational Truth and Analyticity ». In : *Journal of Symbolic Logic* 66.4 (2001), p. 1959–1973 (cf. p. 416).

[126] Volker Halbach. « Disquotationalism and Infinite Conjunctions ». In : *Mind* 108.429 (1999), p. 1–22 (cf. p. 135, 269, 343–344).

[127] Volker Halbach. « How innocent is deflationism ? » In : *Synthese* 126 (2001), p. 167–194 (cf. p. 145, 186, 200, 206, 265, 269–270, 289, 291–292, 322, 330–331).

[128] Volker Halbach. « Reducing compositional to disquotational truth ». In : *The Review of Symbolic Logic* 2.4 (2009), p. 786–798 (cf. p. 416).

[129] Volker Halbach. « Semantics and Deflationism ». Thèse d'habilitation (non publiée). 2001 (cf. p. 9, 25, 206, 269–270).

[130] Volker Halbach et Leon Horsten, éds. *Principles of Truth*. Ontos Verlag, 2003 (cf. p. 269).

[131] Glibert Harman. « Conceptual Role Semantics ». In : *Notre Dame Journal of Formal Logic* 23.2 (1982), p. 242–256 (cf. p. 117).

[132] Richard G. Heck Jr. « The Logical Strength of Compositional Principles ». In : *Notre Dame Journal of Formal Logic* 59.1 (2018), p. 1–33 (cf. p. 300, 303).

[133] Richard G. Heck Jr. « The Strength of Truth Theories ». manuscript non publié. url : http://rgheck.frege.org/philosophy/online_papers.php (cf. p. 300).

[134] Richard G. Heck Jr et Robert May. « Truth in Frege ». In : *Oxford Handbook of Truth*. Sous la dir. de Michael Glanzberg. Oxford University Press, à paraître (cf. p. 8).

[135] Jean van Heijenoort, éd. *From Frege to Gödel : A source book in Mathematical Logic 1879-1931*. 3e. Harvard University Press, 1976 (cf. p. 455).

[136] David Hilbert. « Die Grundlagen der Mathematik ». In : *Abhandlungen aus dem Mathematischen Seminar der Hamburgischen Universität (1928)* 6 (1927). Conférence prononcée en juillet 1927, sur l'invitation du Séminaire de Mathématiques de l'Université de Hambourg. Traduction française in Largeault (1992), p. 145–163 (pour la traduction française) (cf. p. 283).

[137] David Hilbert. « Die Grundlegung des elementaren Zahlentheorie ». In : *Mathematische Annalen* 104 (1931). Traduction anglaise *in* Ewald, 1996, p. 1148-1157, p. 485–494 (cf. p. 211).

[138] David Hilbert. « Neuebegründung der Mathematik. Erste Mitteilung ». In : *Abhandlungen aus dem Mathematischen Seminar der Hamburgischen Universität* 1 (1922). Traduction anglaise *in* Ewald, 1996, p. 157–177 (cf. p. 211).

[139] David Hilbert. « Über das Unendliche ». In : *Mathematische Annalen* 95 (1926). Traduction française *in* (Largeault, 1972). (cf. p. 211, 213–214).

[140] David Hilbert et Paul Bernays. *Grundlagen der Mathematik*. T. II. Berlin : Springer, 1939 (cf. p. 244, 249, 324).

[141] Jaakko Hintikka. *Knowledge and Belief : An Introduction to the Logic of the Two Notions*. Cornell, Ithaca, NY : Cornell University Press, 1962 (cf. p. 393).

[142] Harold Hodes. « On the sense and reference of a logical constant ». In : *The Philosophical Quarterly* 54.214 (2004), p. 134–165 (cf. p. 146, 156).

[143] Wilfrid Hodges. *Model Theory*. T. 42. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1993 (cf. p. 381).

[144] Christopher Hookway. *Quine*. Le Point Philosophique. De Boeck, 1992 (cf. p. 85).

[145] Leon Horsten. « Recension de Cieśliński (2017) ». In : *Notre Dame Philosophical Review* (juin 2018). url : https://ndpr.nd.edu/news/the-epistemic-lightness-of-truth-deflationism-and-its-logic/ (cf. p. 272).

[146] Leon Horsten. « The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth ». In : *The Many Problems of Realism*. Sous la dir. de P. Cortois. T. 3. Studies in the General Philosophy of Science. Tilburg : Tilburg University Press, 1995, p. 173–187 (cf. p. 263).

[147] Leon Horsten. *The Tarskian Turn : Deflationism and Axiomatic Truth*. The MIT Press, 2011 (cf. p. 186, 300, 428).

[148] Leon Horsten et Graham E. Leigh. « Truth is simple ». In : *Mind* 126.501 (jan. 2017), p. 195–232 (cf. p. 380, 402–403, 416, 428).

[149] Paul Horwich. « A defense of minimalism ». In : *Synthese* 126 (2001), p. 149–165 (cf. p. 403).

[150] Paul Horwich. *Meaning*. Oxford, UK : Oxford University Press, 1998 (cf. p. 107–108).

[151] Paul Horwich. *Reflections on Meaning*. Oxford, UK : Oxford University Press, 2005 (cf. p. 107–108).

[152] Paul Horwich. « Three forms of realism ». In : *Synthese* 51 (1982), p. 181–201 (cf. p. 79, 96).

[153] Paul Horwich. *Truth*. 1re éd. Oxford University Press, 1990 (cf. p. 95, 263, 296).

[154] Paul Horwich. *Truth*. 2e éd. Oxford University Press, 1998 (cf. p. 9, 34, 95–108, 119, 133, 135, 175, 192, 235, 253, 267–268, 403).

[155] Paul Horwich. *Truth – Meaning – Reality*. Oxford, UK : Oxford University Press, 2010 (cf. p. 107–108, 403).

[156] Peter Hylton. *Quine*. Routledge, 2007 (cf. p. 84, 90).

[157] Tapani Hyttinen et Gabriel Sandu. « Deflationism and Arithmetical Truth ». In : *Dialectica* 58.3 (2004), p. 413–426 (cf. p. 263–264).

[158] Frank Jackson et Philip Pettit. « Causation in the Philosophy of Mind ». In : *Philosophy and Phenomenological Research* 50.Supplement (1990), p. 195–214 (cf. p. 286–287).

[159] Frank Jackson et Philip Pettit. « Program Explanation : a General Perspective ». In : *Analysis* 50.2 (1990), p. 107–117 (cf. p. 286).

[160] Richard Kaye. *Models of Peano Arithmetic*. Oxford logic guides 15. Oxford University Press, 1991 (cf. p. 77, 311–312, 316–317).

[161] Jeffrey Ketland. « Beth's Theorem and Deflationism — Reply to Bays ». In : *Mind* 118.472 (2009), p. 1075–1079 (cf. p. 170, 301).

[162] Jeffrey Ketland. « Deflationism and Tarski's Paradise ». In : *Mind* 108.429 (1999), p. 69–94 (cf. p. 205, 209, 227, 234, 239, 250, 256, 258, 261, 289, 294, 337–338, 348, 351, 355, 391, 427).

[163] Jeffrey Ketland. « Deflationism and the Gödel Phenomena : Reply to Tennant ». In : *Mind* 114.453 (2005), p. 75–88 (cf. p. 206, 234, 339, 351–356).

[164] Jeffrey Ketland. « Truth, conservativeness and provability : Reply to Cieśliński ». In : *Mind* 119.474 (2010), p. 423–436 (cf. p. 300, 339, 360, 365–366, 377–380, 384–385).

[165] Richard L. Kirkham. *Theories of Truth : A Critical Introduction*. Cambridge, Massachusetts : The MIT Press, 1992 (cf. p. 34, 39).

[166] Roman Kossak et James Henry Schmerl. *The structure of models of Peano arithmetic*. Oxford logic guides 50. Oxford, Clarendon Press, 2006 (cf. p. 312).

[167] Henryk Kotlarski. « Bounded Induction and Satisfaction Classes ». In : *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 32 (1986), p. 531–544 (cf. p. 365).

[168] Henryk Kotlarski. « Full satisfaction Classes : A Survey ». In : *Notre Dame Journal of Formal Logic* 32.4 (1991), p. 573–579 (cf. p. 77, 307, 316).

[169] Henryk Kotlarski, Stanislaw Krajewski et Alistair Lachlan. « Construction of satisfaction classes for non standard models ». In : *Canadian Mathematical Bulletin* 24.3 (1981), p. 283–293 (cf. p. 316–317).

[170] Georg Kreisel. « Abstract : a refinement of ω-consistency ». In : *Journal of Symbolic Logic* 22 (1957), p. 108–109 (cf. p. 246).

[171] Georg Kreisel. « On a Problem of Henkin's ». In : *Indigationes Mathematicae* 15 (1953), p. 405–406. (Cf. p. 368, 375).

[172] Georg Kreisel. « Ordinal logics and the characterization of informal concepts of proof ». In : *Proceedings International Congress of Mathematicians*. 1958, p. 14–21 (cf. p. 211).

[173] Georg Kreisel. « Principles of proof and ordinals implicit in given concepts ». In : *Studies in Logic and the Foundations of Mathematics* 60 (1970), p. 489–516 (cf. p. 211).

[174] Georg Kreisel et Azriel Lévy. « Reflection Principles and their use for establishing the complexity of axiomatic systems ». In : *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 14 (1968), p. 97–142 (cf. p. 343, 365).

[175] Saul Kripke. « Is There a Problem about Substitutional Quantification ? » In : *Truth and meaning : essays in semantics*. Sous la dir. de Gareth Evans et John McDowell. Oxford University Press, 1976. Chap. XIII, p. 325–419 (cf. p. 34).

[176] Saul Kripke. « Naming and necessity ». In : *The Semantics of Natural Language*. Sous la dir. de Donald Davidson et Glibert Harman. Dordrecht : Reidel, 1972, p. 253–355 (cf. p. 111).

[177] Saul Kripke. « Outline of a theory of truth ». In : *Journal of Philosophy* 72.19 (1975), p. 690–716 (cf. p. 353).

[178] Alistair Lachlan. « Full satisfaction classes and recursive saturation ». In : *Canadian Mathematical Bulletin* 24 (1981), p. 295–297 (cf. p. 316).

[179] Jean Largeault, éd. *Intuitionnisme et théorie de la démonstration*. Vrin, 1992 (cf. p. 444).

[180] Jean Largeault, éd. *Logique Mathématique : Textes*. Armand Colin, 1972 (cf. p. 216, 444).

[181] Sandra Laugier. *L'anthropologie logique de Quine : l'apprentissage de l'obvie*. Bibliothèque d'histoire de la philosophie. Vrin, 1992 (cf. p. 85).

[182] Stephen Leeds. « Theories of Reference and Truth ». In : *Erkenntnis* 13 (1978), p. 111–129 (cf. p. 79, 96, 113–114).

[183] Graham E. Leigh. *Conservativity for theories of compositional truth via cut elimination*. arXiv:1308.0168. 2013. url : http://arxiv.org/abs/1308.0168 (cf. p. 317).

[184] Hannes Leitgeb. « What Theories of Truth Should be Like (but Cannot be) ». In : *Philosophy Compass* 2.2 (2007), p. 276–290 (cf. p. 338).

[185] Paolo Leonardi et Marco Santambrogio, éds. *On Quine*. Cambridge University Press, 1995 (cf. p. 452).

[186] Stanislaw Lesniewski. « Grundzüge eines neuen Systems dre Grundlagen der Mathematik ». In : *Fundamenta Mathematicae* 14 (1929), p. 1–81 (cf. p. 42).

[187] Brian Loar. « Conceptual Role and Truth-Conditions ». In : *Notre Dame Journal of Formal Logic* 23 (1982), p. 272–283 (cf. p. 117).

[188] Brian Loar. « Truth Beyond All Verification ». In : *Michael Dummett — Contributions to Philosophy*. Sous la dir. de Barry Taylor. T. 25. Nijhoff International Philosophy Series. Springer, Dordrecht, 1987. Chap. 4, p. 81–116 (cf. p. 96).

[189] Martin Hugo Löb. « Solution of a problem of Leon Henkin ». In : *Journal of Symbolic Logic* 20.2 (1955), p. 115–118 (cf. p. 371).

[190] Kirk Ludwig et Greg Ray. « Semantics for opaque contexts ». In : *Philosophical Perspectives* (1998), p. 141–166 (cf. p. 182–183).

[191] John MacFarlane. « Logical Constants ». In : *The Stanford Encyclopedia of Philosophy*. Sous la dir. d'Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University, 2017. url : https://plato.stanford.edu/archives/win2017/entries/logical-constants/ (cf. p. 426).

[192] Paolo Mancosu. « Explanation in Mathematics ». In : *The Stanford Encyclopedia of Philosophy*. Sous la dir. d'Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University, 2018 (cf. p. 273).

[193] Paolo Mancosu, éd. *From Brouwer to Hilbert : The debate on the foundations of mathematics in the 1920s*. New York et Oxford : Oxford University Press, 1998 (cf. p. 210, 212).

[194] Paolo Mancosu. « Quine et Tarski sur le nominalisme ». In : *Infini, logique, géométrie géométrie*. Vrin, 2015. Chap. V, p. 229–262 (cf. p. 41).

[195] Paolo Mancosu. « Tarski, Neurath et Kokoszyńska sur la conception sémantique de la vérité ». In : *Infini, logique, géométrie*. Vrin, 2015. Chap. III, p. 135–177 (cf. p. 38).

[196] Vann McGee. « In praise of the free Lunch ». In : *Self-Reference*. Sous la dir. de V. F. Hendricks, S. A. Pedersen et T. Bollande. Stanford CSLI Publications, 2006, p. 95–120 (cf. p. 307, 313, 315, 322).

[197] Vann McGee. « Maximal consistent sets of instances of Tarski's schema (T) ». In : *Journal of Philosophical Logic* 21 (1992), p. 235–241 (cf. p. 269).

[198] Brian McLaughlin et Karen Bennett. « Supervenience ». In : *The Stanford Encyclopedia of Philosophy*. Sous la dir. d'Edward N. Zalta. Spring 2018. Metaphysics Research Lab, Stanford University, 2018. URL : https://plato.stanford.edu/archives/spr2018/entries/supervenience/ (cf. p. 110).

[199] Peter Milne. « Classical harmony : Rules of inference and the meaning the logical constants ». In : *Synthese* 100.1 (1994), p. 49–94 (cf. p. 148, 156, 162).

[200] Peter Milne. « On Gödel Sentences and What They Say ». In : *Philosophia Mathematica* 15 (2007), p. 193–226 (cf. p. 238).

[201] Jean-Maurice Monnoyer, éd. *Lire Quine - Logique et Ontologie*. Collection « Lire les philosophes ». Éditions de l'éclat, 2006 (cf. p. 85).

[202] Richard Montague. *Formal philosophy. Selected papers of Richard Montague. Edited and with an introduction by Richmond H. Thomason*. Sous la dir. de Richmond H. Thomason. New Haven : Yale University Press, 1974 (cf. p. 50).

[203] George Edward Moore. « ? » In : *?* (1942) (cf. p. 393).

[204] Yiannis Moschovakis. *Elementary induction on abstract structures*. T. 77. Studies in Logic and the Foundations of Mathematics. Amsterdam : North-Holland Publishing Company, 1974 (cf. p. 62).

[205] Andrzej Mostowski. *Sentences undecidable in formalized arithmetic*. Amsterdam : North-Holland Publishing Company, 1952 (cf. p. 249, 340).

[206] Roman Murawski. « Satisfaction Classes : a Survey ». In : *Euphony and Logos, essays in Honour of Maria Steffen-Batóg and Tadeusz Batóg*. Poznań Studies in the Philosophy of the Sciences and the Humanities 57. Rodopi, 1997 (cf. p. 307).

[207] John Myhill. « Some remarks on the notion of proof ». In : *Journal of Philosophy* 57.14 (1960), p. 461–471 (cf. p. 387).

[208] Ernest Nagel, James R. Newman et Jean-Yves Girard. *Le théorème de Gödel*. Paris, France : Seuil, 1989 (cf. p. 238, 442).

[209] Ilkka Niiniluoto. « Tarskian truth as correspondence — replies to some objections ». In : *Truth and Its Nature (If Any)*. Sous la dir. de Jaroslav Peregrin. Kluwer Academic Publishers, 1999, p. 91–104 (cf. p. 78).

[210] Joseph Ohana, éd. *Langage Vérité et Logique*. Bibliothèque de Philosophie Scientifique. Traduction française de Ayer (1946). 26 rue Racine, Paris : Flammarion, 1956 (cf. p. 433).

[211] David Papineau. « Naturalism ». In : *The Stanford Encyclopedia of Philosophy*. Sous la dir. d'Edward N. Zalta. WInter 2016. Metaphysics Research Lab, Stanford University, 2016. url : https://plato.stanford.edu/archives/win2016/entries/naturalism/ (cf. p. 109–110).

[212] Rohit Parikh. « Existence and Feasibility in Arithmetic ». In : *Journal of Symbolic Logic* 36.3 (1971), p. 494–508 (cf. p. 361, 370).

[213] Rohit Parikh. « Some results on the length of proofs ». In : *Transactions of the American Mathematical Society* 177 (1973), p. 29–36 (cf. p. 370).

[214] Charles Parsons. « Finitism and intuitive knowledge ». In : *The Philosophy of Mathematics today*. Sous la dir. de Matthias Schirn. Oxford University Press, 1998, p. 249–270 (cf. p. 211).

[215] Barbara Hall Partee. « Opacity, coreference and pronouns ». In : *Synthese* 21.3-4 (1970), p. 359–385 (cf. p. 27, 29).

[216] Douglas Patterson. *Alfred Tarski : Philosophy of Language and Logic*. History of Analytic Philosophy. Basingstoke, Hampshire UK : Plagrave Macmillan, 2012 (cf. p. 38–39, 41, 48, 62–63, 83).

[217] Douglas Patterson, éd. *New Essays on Tarski and Philosophy*. Oxford University Press, 2008 (cf. p. 38).

[218] Douglas Patterson. « Tarski on definition, meaning and truth ». In : *The Golden Age of Polish Philosophy*. Sous la dir. de Sandra Lapoint et al. T. 16. Logic, epistemology and the unity of science. Springer, 2009. Chap. 10, p. 155–170 (cf. p. 42).

[219] Karl Popper. *Conjectures and Refutations : The Growth of Scientific Knowledge*. Routledge, 1963 (cf. p. 78).

[220] Karl Popper. *Logic of Scientific Discovery*. Routledge Classics. New York : Routledge, 2005 (cf. p. 39).

[221] Dag Prawitz. *Natural Deduction, A Proof-theoretical Study*. Stokholm : Almquist et Wiksell, 1965 (cf. p. 139–140).

[222] Arthur Norman Prior. « The runabout inference-ticket ». In : *Analysis* 21.2 (1960), p. 38–39 (cf. p. 140).

[223] Hilary Putnam. *Meaning and the Moral Sciences*. Londres : Routledge & Kegan Paul Ltd, 1978 (cf. p. 114, 285).

[224] Hilary Putnam. *Philosophy of logic*. Harper & Row New York, 1971 (cf. p. 220).

[225] Willard Van Orman Quine. « Carnap and Logical Truth ». In : *The Ways of Paradox and other essays*. Harvard University Press, 1954 (cf. p. 86).

[226] Willard Van Orman Quine. *From a Logical Point of View*. Boston, Massachusetts : Harvard University Press, 1980 (cf. p. 451).

[227] Willard Van Orman Quine. *La Poursuite de la vérité*. Traduit de l'anglais par Maurice Clavelin. Paris, France : Seuil, l'ordre philosophique, 1993 (cf. p. 451).

[228] Willard Van Orman Quine. *Le Mot et la Chose (traduction de* Quine, 1960 *par Joseph Dopp & Paul Gochet)*. Champs. Flammarion, 1999 (cf. p. 78, 179, 181, 452).

[229] Willard Van Orman Quine. *Les voix du paradoxe et autres essais*. Sous la dir. de Serge Bozon et Sabine Plaud. Vrin, 2011 (cf. p. 452).

[230] Willard Van Orman Quine. *Mathematical Logic*. Boston, Massachusetts : Harvard University Press, 1940 (cf. p. 175, 177–178).

[231] Willard Van Orman Quine. « Notes on the Theory of Reference ». In : *From a Logical Point of View*. Harvard University Press, 1953. Chap. 7 (cf. p. 85).

[232] Willard Van Orman Quine. « On what there is ». In : *The Review of Metaphysics* 2.5 (1948). Repris *in* Quine, 1980, p. 21–48 (cf. p. 221).

[233] Willard Van Orman Quine. *Philosophie de la logique*. Traduit de l'anglais par Jean Largeault, présentation par Denis Bonnay et Sandra Laugier. Traduction française de (Quine, 1970). Aubier, 2008 (cf. p. 91–92, 451).

[234] Willard Van Orman Quine. *Philosophy of Logic*. Traduction française par Jean Largeault Quine (2008). Prentice Hall, 1970 (cf. p. 85–86, 88–89, 96, 100, 105, 113, 119, 133, 235, 451).

[235] Willard Van Orman Quine. *Pursuit of Truth*. Traduction française Maurice Clavelin (Quine, 1993). Cambridge, Massachusetts : Harvard University Press, 1990 (cf. p. 84–86, 94, 133, 189).

[236] Willard Van Orman Quine. *Quiddités*. Traduit de l'anglais par Dominique Goy-Blanquet et Thierry Marchaisse. Paris, France : Seuil, l'ordre philosophique, 1992 (cf. p. 452).

[237] Willard Van Orman Quine. *Quiddities - An Intermmitently Philosophical Dictionnary*. Traduction françaies Quine (1992). Harvard University Press, 1987 (cf. p. 84, 86).

[238] Willard Van Orman Quine. « Reactions ». In : Leonardi *et* Santambrogio, *1995*. Sous la dir. de Paolo Leonardi et Marco Santambrogio. Cambridge University Press, 1995. Chap. 20, p. 347–361 (cf. p. 90).

[239] Willard Van Orman Quine. « Truth by Convention ». In : *Philosophical Essays for A. N. Whitehead*. Sous la dir. d'O. H. Lee. Traduction française in Quine (2011). New York : Longmans, 1935 (cf. p. 86).

[240] Willard Van Orman Quine. « Two Dogmas of Empiricism ». In : *Philosophical Review* (1951) (cf. p. 86).

[241] Willard Van Orman Quine. *Word and Object*. Traduction française Quine, 1999. The MIT Press, 1960 (cf. p. 86, 88–90, 451).

[242] Panu Raatikainen. « Hilbert's program revisited ». In : *Synthese* 137.1-2 (2003), p. 157–177 (cf. p. 210–211, 215, 219).

[243] Frank Plumpton Ramsey. « Facts and propositions ». In : *Proceedings of the Aristotelian Society, Supplementary Volumes* 7 (1927). Traduction française in Ramsey (2003, p. 213–228) (cf. p. 15–16, 18–20, 22, 96, 99).

[244] Frank Plumpton Ramsey. *Logique, philosophie et probabilités*. Sous la dir. de Pascal Engel et Mathieu Marion. Vrin, 2003 (cf. p. 452).

[245] Frank Plumpton Ramsey. *On Truth*. Sous la dir. de Nicholas Rescher et Ulrich Majer. T. 16. Episteme. Original manuscript materials (1927-1929) from the Ramsey Collection at the University of Pittsburgh. Kluwer Academic Publishers, 1991 (cf. p. 15–22, 26, 96, 101).

[246] Greg Ray. « On the matter of essential richness ». In : *Journal of Philosophical Logic* 34 (2005), p. 433–457 (cf. p. 54, 60, 62–63, 67, 251).

[247] Stephen Read. « Harmony and autonomy in classical logic ». In : *Journal of Philosophical Logic* 29 (2000), p. 123–154 (cf. p. 148–149, 152).

[248] Stephen Read. *Relevant Logic: a philosophical examination of inference*. Oxford, UK : Basil Blackwell, 1988 (cf. p. 148–149).

[249] Michael Resnik. « How nominalist is Hartry Field's nominalism ? » In : *Philosophical Studies* 47 (1985), p. 163–181 (cf. p. 222).

[250] Mark Richard. « Quotation, Grammar, and Opacity ». In : *Linguistics and Philosophy* 9.3 (1986), p. 383–403 (cf. p. 182–183).

[251] François Rivenc. « Ce que Ramsey a vraiment dit, ou la théorie prophrastique de la vérité ». In : *Philosophie* 57 (mar. 1998), p. 16–50 (cf. p. 14, 17–18, 25, 28, 31, 34).

[252] François Rivenc. « Contre la déflation de la vérité ». In : *Dialectica* 58.4 (2004), p. 517–528 (cf. p. 187, 194).

[253] François Rivenc. *Lecture de Quine*. Cahiers de Logique et d'Epistémologie. College Publications, 2008 (cf. p. 85).

[254] François Rivenc. « Théories de la vérité et sémantique des conditions de vérité : le projet de Tarski ». In : *Les Études philosophiques* 3 (1996), p. 382–402 (cf. p. 88).

[255] Artur Rojszczak. *From the act of judging to the sentence : the problem of truth bearers from Bolzano to Tarski*. Sous la dir. de Jan Woleński. T. 328. Synthese Library. Springer, 2005 (cf. p. 41).

[256] John Barkley Rosser. « Gödel theorems for non-constructive logic ». In : *Journal of Symbolic Logic* 2.3 (1937), p. 129–137 (cf. p. 340).

[257] Philippe de Rouilhan. « Note sur Popper lecteur de Tarski ». In : *Philosophia Scientiae* 11.1 (2007), p. 131–148 (cf. p. 78).

[258] Philippe de Rouilhan. « Tarski et l'universalité de la logique ». In : *Le formalisme en question*. Sous la dir. de Fréderic Nef et Denis Vernant. Paris, France : Vrin, 1998 (cf. p. 37, 54–55, 58–60, 62, 251).

[259] Richard Schantz. « Was Tarski a deflationist ? » In : *Logic and Logical Philosophy* 6 (1998), p. 157–172 (cf. p. 78).

[260] Paul Arthur Schilp, éd. *The Philosophy of Rudolf Carnap*. The Library of Living Philosophers. La Sale, Illinois : Open Court, 1963 (cf. p. 39).

[261] Stewart Shapiro. « Conservativeness and incompleteness ». In : *Journal of Philosophy* 80 (1983), p. 521–531 (cf. p. 222–224, 226–227, 240).

[262] Stewart Shapiro. « Deflation and Conservation ». In : *Principles of Truth*. Sous la dir. de Volker Halbach et Leon Horsten. Ontos Verlag, 2003, p. 103–128 (cf. p. 264, 290, 294, 322–323, 331, 338–339).

[263] Stewart Shapiro. « Induction and indefinite extensibility : the Gödel sentence is true, but did someone change the subject ? » In : *Mind* 107.427 (1998), p. 597–624 (cf. p. 206, 250).

[264] Stewart Shapiro. « Proof and Truth : Through Thick and Thin ». In : *Journal of Philosophy* 95.10 (1998), p. 493–521 (cf. p. 205, 209, 227–228, 230, 233, 237–239, 258, 261, 289, 294, 296, 318–319, 337–338, 348, 351, 391, 427).

[265] Stewart Shapiro. « Systems between First-order and Second-order logics ». In : *Handbook of Philosophical Logic*. Sous la dir. de Dov Gabbay et Franz Guenther. 2e éd. T. 1. Kluwer Academic Publishers, 2001, p. 131–187 (cf. p. 34).

[266] Gila Sher. « In search of a substantive theory of truth ». In : *Journal of Philosophy* 101.1 (2004), p. 5–36 (cf. p. 78).

[267] Gila Sher. « On the possibility of a substantive theory of truth ». In : *Synthese* 117.1 (1999), p. 133–172 (cf. p. 78).

[268] Peter Simons. « Lesniewski and Ontological Commitment ». In : *Stanislaw Lesniewski aujourd'hui*. Sous la dir. de Denis Miéville et Denis Vernant. Vrin, 1996, p. 103–120 (cf. p. 41).

[269] Peter Simons. « Reasoning on a Tight Budget: Lesniewski's Nominalistic Metalogic ». In : *Erkenntnis* 56.1 (2002), p. 99–122 (cf. p. 41).

[270] Peter Simons. « Truth on a Tight Budget : Tarski and Nominalism ». In : *New Essays on Tarski and Philosophy*. Sous la dir. de Douglas Patterson. Oxford University Press, 2008. Chap. 14 (cf. p. 41).

[271] Peter Simons et Jan Woleński. « De Veritate : Austro-Polish Contributions to the Theory of Truth from Brentano to Tarski ». In : *The Vienna Circle and the Lvov-Warsaw School*. T. 38. Nijhoff International Philosophy Series. Springer, 1989, p. 391–442 (cf. p. 38).

[272] Stephen George Simpson. « Partial realizations of Hilbert's program ». In : *The Journal of Symbolic Logic* 53.2 (1988), p. 349–363 (cf. p. 219).

[273] Stephen George Simpson. *Subsystems of second order arithmetic*. 2e éd. Perspectives in mathematical logic. Berlin : Springer, 2009 (cf. p. 213, 219, 323–324, 326–327, 330).

[274] Thoralf Skolem. « Begründung der elementaren Arithmetik durch die rekurrierende Denkweise ohne Anwendung scheinbarer Veränderlichen mit unendlichem Ausdehnungsbereich ». In : *Videnskapsselskapets skrifter, I. Matematisk-naturvidenskabelig klasse* 6 (1923). Traduction anglaise : *The foundations of elementary arithmetic established by means of the recursive mode of thought, without the use of apparent variables ranging over infinite domains, in* Heijenoort, 1976 (cf. p. 213).

[275] Peter Smith. *An Introduction to Gödel's Theorems*. 2e éd. Cambridge University Press, 2013 (cf. p. 247).

[276] Stuart Smith. « Non-standard syntax and semantics and full satisfaction classes ». Thèse de doct. New Haven, Connecticut, USA : Yale University, 1984 (cf. p. 316).

[277] Stuart Smith. « Nonstandard definability ». In : *Annals of Pure and Applied Logic* 42 (1989), p. 21–43 (cf. p. 316).

[278] Craig Smoryński. « The development of self-reference : Löb's theorem ». In : *Perspectives on the History of Mathematical Logic*. Sous la dir. de Thomas Drucker. Modern Birkhäuser Classics. Birkhäuser, 1991, p. 110–133 (cf. p. 371).

[279] Craig Smoryński. « The Incompleteness Theorems ». In : *Handbook of Mathematical Logic*. Sous la dir. de Jon Barwise. T. 90. Studies in Logic and the Foundations of Mathematics. North-Holland, 1977. Chap. D1, p. 821–865 (cf. p. 216, 371).

[280] Daniel Stoljar. « Physicalism ». In : *The Stanford Encyclopedia of Philosophy*. Sous la dir. d'Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University, 2017. url : https://plato.stanford.edu/archives/win2017/entries/physicalism/ (cf. p. 109–110).

[281] Daniel Stoljar et Nic Damnjanovic. « The Deflationary Theory of Truth ». In : *The Stanford Encyclopedia of Philosophy*. Sous la dir. d'Edward N. Zalta. Fall 2014. Metaphysics Research Lab, Stanford University, 2014 (cf. p. 9).

[282] Peter Frederick Strawson. « A Problem About Truth—A Reply to Warnock ». In : *Truth*. Sous la dir. de George Pitcher. Englewood Cliffs, NJ : Prentice-Hall, 1964 (cf. p. 96).

[283] Peter Frederick Strawson. « Truth ». In : *Proceedings of the Aristotelian Society, Supplementary Volumes* 24 (1950), p. 129–56 (cf. p. 96).

[284] Andrea Strollo. « Deflationism and the invisible power of truth ». In : *Dialectica* 67.4 (2013), p. 521–543. doi : 10.1111/1746-8361.12044 (cf. p. 300, 307–308, 310, 313, 315, 318, 320, 322).

[285] William Walker Tait. « Finitism ». In : *Journal of Philosophy* 78 (1981), p. 24–546 (cf. p. 211–212).

[286] William Walker Tait. « Remarks on finitism ». In : *Reflecions on the Foundations of Mathematics. Essays in Honor of Solomon Feferman*. Sous la dir. de William Sieg, Richard Sommer et Carolyn Talcott. T. 15. Association for Symbolic Logic, LNL, 2002 (cf. p. 211).

[287] Gaisi Takeuti. *Proof Theory*. 2e éd. T. 81. Studies in Logic and the Foundations of Mathematics. Amsterdam : North-Holland, 1987 (cf. p. 329).

[288] Alfred Tarski. « Der Wahrheitsbegriff in den formalisierten Sprachen ». In : *Studia Philosophica (1936) (tiré à part daté de 1935)* 1 (1935). Traduit en français sous le titre « *Le concept de vérité dans les langages formalisés* » *in* Tarski, 1972, p. 261–405 (cf. p. 37–42, 46, 48–50, 54–65, 68–69, 72–76, 78–80, 82, 87, 173, 177–179, 234, 240, 249, 251, 254, 337–338, 340, 403).

[289] Alfred Tarski. « Der Wahrheitsbegriff in den Sprachen des deduktiven Disziplinen ». In : *Anzeiger der Osterreichischen Akademie der Wissenschaften, Mathematisch-Naturwissenschaftliche Klasse* 69 (1932), p. 23–25 (cf. p. 37).

[290] Alfred Tarski. *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. Sous la dir. de J. H. Woodger. 1re éd. Oxford University Press, 1956 (cf. p. 37).

[291] Alfred Tarski. *Logique, sémantique, métamathématiques 1923-1944*. T. I. Traduction française sous la direction de Gilles-Gaston Granger. Armand Colin, 1972 (cf. p. 37–38, 456).

[292] Alfred Tarski. *Logique, sémantique, métamathématiques 1923-1944*. T. II. Traduction française sous la direction de Gilles-Gaston Granger. Armand Colin, 1974 (cf. p. 38, 457).

[293] Alfred Tarski. « O pojęciu prawdy w odniesieniu do sformalizowanych nauk dedukcyjny ». In : *Ruch Filozoficzny* 12 (1930-31), p. 210–211 (cf. p. 37).

[294] Alfred Tarski. « Pojęcie prawdy w jezykach nauk dedukcyjnych (Le concept de vérité dans le langage des sciences déductives) ». In : *Towarzystwo Naukowe Warszawskie* (1933) (cf. p. 37, 54–56, 61–62).

[295] Alfred Tarski. « Some Methodological Investigations on the Definability of Concepts ». In : *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. Sous la dir. de J. H. Woodger. (Traduction française *in* Tarski (1974, p. 23–46)). Oxford, Clarendon Press, 1956. Chap. 10, p. 296–319 (cf. p. 63).

[296] Alfred Tarski. « The semantic conception of truth and the foundations of semantics ». In : *Philosophy and Phenomenological Research* 4 (1944). Traduction française *in* Tarski (1974, p. 267–305) (cf. p. 37–38, 40–44, 46–54, 61, 63–65, 79–82, 436).

[297] Alfred Tarski. « Truth and Proof ». In : *Scientific American* 220 (1969), p. 63–77 (cf. p. 37, 79, 238).

[298] Neil Tennant. *Anti-Realism and Logic : Truth as Eternal*. Clarendon Library of Logic et Philosophy, Oxford University Press, 1987 (cf. p. 162).

[299] Neil Tennant. « Deflationism and the Gödel Phenomena ». In : *Mind* 111.443 (2002), p. 551–582 (cf. p. 206, 238, 264, 276, 337–339, 346–349, 351–352, 354–356, 358–359, 362, 364–365, 370, 377, 380, 383, 385, 403).

[300] Neil Tennant. « Deflationism and the Gödel Phenomena : Reply to Cieśliński ». In : *Mind* 119.474 (2010), p. 437–450 (cf. p. 206, 264, 268, 270, 276, 339, 360, 365, 377, 403).

[301] Neil Tennant. « Deflationism and the Gödel Phenomena : Reply to Ketland ». In : *Mind* 114.453 (2005), p. 89–96 (cf. p. 206, 264, 339, 351, 356–360, 362, 364–365, 369–370, 376–377, 403).

[302] Neil Tennant. *The Taming of the True*. Clarendon Press, Oxford, 1997 (cf. p. 162).

[303] A. S. Troelstra et H. Schwichtenberg. *Basic Proof Theory*. Seconde. T. 43. Cambridge Tracts in Theoretical Computer Science. Cambridge, UK : Cambridge University Press, 2001 (cf. p. 151).

[304] Alan Mathison Turing. « Systems of logic based on ordinals ». In : *Proceedings of the London Mathematical Society* 2 (1939), p. 161–228 (cf. p. 340).

[305] Alasdair Urquhart. « The logic of physical theory ». In : *Physicalism in Mathematics*. Sous la dir. d'A. D. Irvine. Dordrecht, Kluwer, 1990, p. 145–154 (cf. p. 222–223).

[306] Bas Cornelis Van Fraassen. *The Scientific Image*. Oxford University Press, 1980 (cf. p. 387–388).

[307] Frédéric Vuissoz. *La conception sémantique de la vérité - Logique et philosophie chez Alfred Tarski*. T. 12. Travaux de logique. Centre de Recherches Sémiologiques, Université de Neufchâtel, déc. 1998 (cf. p. 38).

[308] Paul Weingartner. « Tarski's Truth Condition Revisited ». In : *Alfred Tarski and the Vienna Circle : Austro-Polish Connections in Logical Empiricism*. Sous la dir. de Jan Woleński et Eckehart Köhler. T. 6. Vienna Circle Institute Yearbook. Kluwer Academic Publishers, 1999, p. 193–201 (cf. p. 78).

[309] Daniel Whiting. « Conceptual Role Semantics ». In : *The Internet Encyclopedia of Philosophy*. Sous la dir. de James Fieser et Bradley Dowden. ISSN 2161-0002, 2018. url : https://www.iep.utm.edu/conc-rol/ (cf. p. 117).

[310] Christopher John Fards Williams. *What is Truth ?* Cambridge University Press, 1976 (cf. p. 24–25).

[311] Michael Williams. « Do We (Epistemologists) Need A Theory of Truth ». In : *Philosophical Topics* 14.1 (1986), p. 223–242 (cf. p. 96).

[312] Ludwig Wittgenstein. *Philosophical Investigations*. Traduction anglaise par G. E. M. Anscombe. Oxford : Blackwell, 1953 (cf. p. 96).

[313] Ludwig Wittgenstein. *Recherches Philosophiques*. Sous la dir. d'Élisabeth Rigal. Tel. Gallimard, 2004 (cf. p. 393).

[314] Ludwig Wittgenstein. *Tractatus Logico-philosophicus*. Traduction française et édition de Gilles-Gaston Granger, Gallimard, 1993. Routledge & Kegan Paul Ltd, 1922 (cf. p. 96, 156).

[315] Jan Woleński. « The Rise and Development of Logical Semantics in Poland ». In : *The Golden Age of Polish Philosophy*. Sous la dir. de Sandra Lapoint et al. T. 16. Logic, epistemology and the unity of science. Springer, 2009. Chap. 3, p. 43–59 (cf. p. 42).

[316] Crispin Wright. « Intuition, entitlement and the epistemology of logical laws ». In : *Dialectica* 58.1 (2004), p. 155–175 (cf. p. 392).

[317] Crispin Wright. « Warrant for nothing (and foundations for free) ? » In : *Proceedings of the Aristotelian Society, Supplementary Volumes* 78.167-212 (2004) (cf. p. 392).

[318] Richard Zach. « Hilbert's Program ». In : *The Stanford Encyclopedia of Philosophy (Spring 2009 Edition), Edward N. Zalta (ed.)* (2009). URL : http://plato.stanford.edu/archives/spr2009/entries/hilbert-program/ (cf. p. 210, 213).

Recherches sur le déflationnisme aléthique contemporain

Résumé : « Qu'est-ce que la vérité ? » À cette question, les déflationnistes aléthiques contemporains proposent une réponse originale : la propriété de vérité ne serait qu'un simple outil de décitation, indispensable pour formuler certaines généralisations mais dénué de tout pouvoir explicatif propre. Selon eux, elle ne jouerait donc pas de rôle important dans notre activité scientifique. L'objectif de cette thèse est d'évaluer la solidité de la position déflationniste en la confrontant à divers arguments avancés contre ce type de conceptions de la vérité. Après avoir précisé les doctrines centrales du déflationnisme actuel, notre travail se poursuit en deux parties, que l'on peut voir comme deux tentatives complémentaires de fournir un cadre méthodologique permettant d'examiner précisément les théories déflationnistes de la vérité. Dans un premier temps, nous analysons la thèse, souvent attribuée aux déflationnistes, selon laquelle le prédicat de vérité serait une sorte de notion logique. Dans un second temps nous examinons un célèbre argument anti-déflationniste appelé « argument de la conservativité ». Au final, si le déflationnisme ne nous paraît pas totalement désarmé face aux critiques dont il a fait l'objet, notre travail a néanmoins permis de montrer que certaines réponses majeures avancées pour sa défense ne sont plus tenables.

Inquiries into contemporary alethic deflationism

Abstract : « What is truth ? » Contemporary alethic deflationists offer an original answer to this question : the truth predicate is nothing more than a simple disquotation device, allowing us to assert infinite lots of sentences but devoid of any proper explanatory power. It thus has no substantial role to play in our scientific theorizing. The purpose of this dissertation is to assess the soundness of modern deflationist positions by confronting them with several anti-deflationist arguments. After carefully delineating basic tenets of current deflationism, we pursue our investigations into two parts, which can be considered as two complementary ways to set up a precise methodological framework in order to rigorously assess deflationist theories of truth. Firstly, we examine the view, often attributed to deflationists, according to which the truth predicate is akin to a logical notion. Secondly, we analyse a famous argument raised against deflationary conceptions of truth, known as the « conservativity argument ». In the end, although we rebut some prominent deflationist answers, we believe there might still be a way out for deflationism.

Discipline : Philosophie

Mots-clés : Vérité ; Déflationnisme ; Logique ; Épistémologie ; Philosophie de la logique ; Philosophie des mathématiques ; Théories axiomatiques ; Rationalité ; Incomplétude.

Équipe d'accueil : Institut d'Histoire et de Philosophie des Sciences et des Techniques (UMR 8590), 13 rue du Four, 75006 Paris

École doctorale : École doctorale de Philosophie de l'Université Paris 1 Panthéon-Sorbonne (ED 280), 1 rue d'Ulm, 75005 Paris