# L'interprétation double de la probabilité et les problèmes d'auto-localisation

Laurent Delabre

**Université Paris 1 Panthéon-Sorbonne**

**École doctorale de philosophie**

Thèse pour obtenir le grade de docteur en philosophie,
présentée et soutenue publiquement par Laurent DELABRE

# L'interprétation double de la probabilité et les problèmes d'auto-localisation

Thèse dirigée par M. le professeur Pierre WAGNER

à l'Institut d'histoire et de philosophie
des sciences et des techniques (IHPST, UMR 8590),

et soutenue le 8 février 2019

**Jury** :

- Mme Anouk BARBEROUSSE, professeure à Sorbonne Université (rapporteuse)
- M. Mikaël COZIC, maître de conférences HDR à l'université Paris-Est Créteil
- M. Thierry MARTIN, professeur émérite à l'université de Franche-Comté (rapporteur)
- M. Pierre WAGNER, professeur à l'université Paris 1 Panthéon-Sorbonne
- M. Bernard WALLISER, chercheur à Paris Sciences Économiques

## Remerciements

Je remercie les chercheurs de l'Institut d'histoire et de philosophie des sciences et des techniques (IHPST, UMR 8590) pour leur dévouement et leur accueil, et pour tout ce que j'ai appris à leur contact.

Je remercie vivement Paul Franceschi, de l'université de Corse, qui m'a initié à la philosophie analytique par la voie des paradoxes.

Je remercie chaleureusement Léo Gerville-Réache, de l'université de Bordeaux, pour toutes nos discussions enflammées enrichissantes.

## Sommaire

# Introduction

Dans *L'émergence de la probabilité*, Ian Hacking montre que le concept de probabilité est né avec une profonde, puissante et tumultueuse ambivalence qui résiste aux efforts des analystes et des conciliateurs :

> J'ai tenté d'expliquer comment, dès l'origine, s'est installée une tension constante entre la face dite objective et la face dite subjective de la probabilité. Je déteste ces deux épithètes, « objective » et « subjective », mais la tension est bien réelle. Une analyse soigneuse – exercice auquel un philosophe analytique comme moi est assez bien rodé – ne permet pas de s'en affranchir. Cette tension crée des difficultés philosophiques sans solution. Elle est irrévocable : je voulais comprendre pourquoi.[1]

« Tension constante », « déteste », « difficultés », « sans solution », « irrévocable » : l'auteur semble apporter son témoignage tragique, peut-être volontairement exagéré. Un amateur de paradoxes probabilistes (Hacking lui-même s'est notamment penché sur le paradoxe de Goodman dans *Le plus pur nominalisme*) pourrait reconnaitre dans ces quelques lignes une expérience désagréable liée à sa passion.

Pourquoi Hacking déteste-t-il « objectif » et « subjectif » ? D'abord parce que la gradation dans l'objectivité (ou la subjectivité) d'une mesure ou d'une évaluation empêche de penser une dualité : à ce niveau d'analyse, une probabilité ne se range pas dans une case parmi deux seulement, elle

[1] Hacking (2002), p. 15-16.

n'est pas, ou bien existante en soi et déterminée par une expérience et un calcul infaillible, ou bien au contraire dépendante du sujet qui la pense et donc variable, mais elle est *plus ou moins* dépendante. Certes, des mathématiciens et philosophes comme Poincaré ont très bien utilisé les qualificatifs « objectif » et « subjectif » pour rendre compte d'une double interprétation d'un unique concept de probabilité, pour éviter de penser frontalement une probabilité qui serait une propriété des étants du monde, détachée d'une probabilité qui concernerait l'état de nos connaissances. Mais quand nous nous tournerons vers Carnap, autre philosophe influent, nous comprendrons que la paire « ontique »/« épistémique » est plus appropriée : Carnap insiste en effet, au moins autour de l'année 1950, sur l'objectivité des deux concepts de probabilité qu'il distingue, sur l'objectivité des deux discours qui les emploient[2]. Les deux contraires « objectif » et « subjectif » ne semblent donc pas capturer la dualité.

Cette dualité de la probabilité est, selon Hacking, antérieure aux travaux de Pascal et de Fermat, puisque déjà présente dans le mot « possible ». Pensons au contraste entre « possible que » et « possible de » : « il est *possible qu'*Usain Bolt ait couru le 100 mètres en moins de 9 secondes 50 à l'entrainement » signifie que, *pour autant qu'on sache*, on ne peut pas exclure qu'il ait réussi cet exploit ; mais « il est *possible de* courir le 100 mètres en moins de 9 secondes 60 quand on s'appelle Usain Bolt » signifie que le champion est *physiquement capable* d'un tel exploit. Il y a une possibilité épistémique dans le premier cas, physique dans le second. Malgré l'ambiguïté, le vieux mot « équipossibilité » fut employé pendant plus de deux siècles par les penseurs de la probabilité, ce qui fait dire à Hacking que c'est *en vertu* de l'ambiguïté que « possibilité » put efficacement définir la mystérieuse « probabilité » : cette définition

[2] C'est à Barberousse (2000), p. 18-20, que nous devons ces considérations à la fois sur Poincaré et Carnap.

équivoque n'est pas monstrueuse, c'est un trait essentiel du développement du concept, c'est même un avantage[3].

Il ne fait aucun doute qu'en dehors de « subjectif » ou « subjectiviste », le terme « épistémique » est aujourd'hui presque unanimement employé pour qualifier la probabilité perçue comme l'évaluation de l'incertitude, comme l'intensité quantifiable de la confiance accordée par un sujet à une proposition qu'il est permis de peser autrement que dans la binarité classique vrai/faux. Même lorsque cette probabilité est précisément un degré de croyance (à distinguer par exemple du degré de confirmation d'une hypothèse), le mot « doxastique » est abandonné au profit d'« épistémique » qui suggère que la croyance en question est candidate au rang de connaissance. « Épistémique » aura donc partout notre préférence. Le mot qu'il convient de justifier est « ontique ». Pourquoi « probabilité physique », formule plutôt répandue, n'aura pas nos faveurs ? Pour une raison assez similaire à celle de notre refus de « probabilité aléatoire », expression plus rare dans la littérature mais notamment trouvée sous la plume de Hacking. D'abord, la langue anglaise a la chance de posséder « *random* » et « *aleatory* », que nous traduisons indistinctement par « aléatoire ». Avec un seul mot, nous ne pouvons pas facilement distinguer « qui arrive par hasard » et « propre à ce qui arrive par hasard » ; en d'autres occasions cela serait sans importance, mais nous voulons aujourd'hui la liberté d'employer le mot « aléatoire » sans lui donner le sens de « non épistémique ». Ensuite, nous pensons qu'une probabilité n'est pas toujours liée à ce qui arrive par hasard, c'est-à-dire à un événement imprévisible du monde, et nous verrons plus tard comment elle peut être une proportion relative d'événements nécessaires et déterminés, ou de combinaisons de représentations mentales d'événements. Le mot « physique » est lui aussi inadéquat ; en outre, D'Alembert, personnage

[3] Hacking (2002), p. 174.

important de notre thèse, l'utilise, pratiquement à contresens relativement à notre interprétation spontanée, pour qualifier, non pas une mesure donnée dans une expérience du monde, mais un nombre que l'on s'attend à découvrir dans une nature en partie inconnue. Nous estimons en revanche que les épithètes « ontologique » et, mieux encore, « ontique », de plus en plus utilisées par les philosophes[4], évitent la plupart des reproches, que la probabilité par elles qualifiée soit une propension, une proportion, une fréquence, ou une autre propriété des objets de la nature *ou de leurs représentations* : nous considérons que les objets physiques ne sont pas seuls à avoir de l'être.

« Ontique », « épistémique »… Soit. Mais nous n'évincerons pas tout usage des termes « objectif » et « subjectif », qui ne doivent pas être compris comme de mauvais synonymes des deux autres. Remarquer l'objectivité d'un calcul ou la subjectivité d'une estimation, c'est se situer à un niveau de l'analyse de l'aléatoire qui complète le niveau d'analyse de l'essence de la probabilité. Les quatre adjectifs permettent des combinaisons intéressantes.

La tension ontique-épistémique engendre des difficultés, écrivait Hacking. Elle est une surveillance, plus ou moins soutenue selon les époques et les écoles, du visage ontique sur le visage épistémique et inversement, non seulement au sein de la théorie, mais dans le calcul : par exemple, s'il s'écarte de chances objectives, un degré de croyance en la réalisation d'un événement est suspect. Les théoriciens qui veulent maintenir deux interprétations ne parviennent pas à les mêler dans une convaincante et infaillible unité de discours ; moins nombreux, les « objectivistes » radicaux et les « subjectivistes » radicaux éprouvent quant à eux le manque cruel de la moitié abandonnée. Des expressions comme « une interprétation double » ou « la probabilité une et deux » sont

[4] Cf. Cozic et Walliser (2012) pour « ontologique » et Armatte (2011) pour « ontique ».

artificielles, elles n'ont que très peu de sens, elles ne servent qu'à désigner le problème. Nous aimerions évidemment dévoiler tout le sens : il ne se construit que dans la longue recherche méthodique et, peut-être, la découverte d'une porte vers une fin heureuse. Hacking a choisi d'aller aux origines de la probabilité, grâce à son « archéologie » qui pénètre des terrains peu explorés par les historiens de la philosophie et des mathématiques ; nous choisissons une autre voie, en affrontant des paradoxes ardus qui opposent deux interprétations de la probabilité. Nous croyons en effet que la tension ontique-épistémique est le Paradoxe suprême de la théorie des probabilités, un tourment, un diable pour l'épistémologie. Il semble invulnérable, mais il montre des faiblesses quand il s'incarne dans un paradoxe particulier. C'est là qu'il faut frapper.

**De quel paradoxe en particulier parlons-nous ?**

Arnold Zuboff est un philosophe américain peu connu, qui publie des articles depuis bientôt quarante ans mais n'a obtenu son doctorat à Princeton qu'en 2009. En 1986, concentré sur son thème favori, l'identité personnelle, il rédige un manuscrit qu'il montre dans un premier temps à l'épistémologue et métaphysicien Peter Unger, qui en fait parvenir une copie à Robert Stalnaker, un spécialiste de la sémantique des mondes possibles. Celui-ci est intrigué par des expériences de pensée décrites par Zuboff, où des individus sont drogués, endormis pendant un très long moment et réveillés au hasard un certain nombre de fois. En 1990, lors de la publication de l'article de Zuboff, « *One Self: The Logic of Experience* », les expériences de sommeil prolongé sont peu remarquées. Pendant quelques années, on ne parle plus d'un dormeur qui cherche à se localiser dans le temps[5].

[5] Nous reconstituons ce récit des origines de *Sleeping Beauty* à partir, notamment, des témoignages de Zuboff et d'Elga. Cf. Elga (2000), note 1 ; Zuboff (2009), p. 2.

Un document de travail de Michele Piccione et Ariel Rubinstein, apparu en 1994, est publié trois ans plus tard sous le titre « *On the Interpretation of Decision Problems with Imperfect Recall* ». Les deux économistes passionnés par la théorie des jeux y présentent surtout le problème du Conducteur distrait (*Absent-Minded Driver*), où un automobiliste à la mémoire défaillante doit trouver la stratégie optimale pour espérer retourner chez lui : il sait qu'il doit prendre la deuxième sortie de l'autoroute, mais sait aussi qu'arrivé au niveau d'une sortie il sera incapable de reconnaitre s'il s'agit de la première ou de la seconde. En 1999, le philosophe Adam Elga utilise la structure du Conducteur distrait (l'agent rationnel sait qu'il sera au niveau d'une sortie à l'instant $t_1$, mais pourra être ou ne pas être au niveau d'une sortie à l'instant $t_2$) pour fabriquer un problème probabiliste. Il est aidé dans cette tâche par Robert Stalnaker qui se souvient des exemples de Zuboff et donne au nouveau problème le nom de *Sleeping Beauty* en référence au célèbre conte. Elga présente le Conducteur distrait et la Belle au bois dormant lors d'une conférence en présence de Jamie Dreier, lequel va donner à ces problèmes une audience un peu plus large en les énonçant sur un forum de discussion du nom de rec.puzzles[6] : c'est donc via Internet que des étudiants, particulièrement fascinés par le paradoxe probabiliste engendré par la Belle au bois dormant, défendent pour la première fois l'une ou l'autre de ses solutions en rivalisant d'astuce. En 2000, Elga publie un article sur la Belle au bois dormant dans la revue *Analysis* et y propose une solution. L'année suivante, David Lewis, avocat de la solution adverse, lui répond ; il décède en octobre 2001 sans avoir pu développer davantage son point de vue et se défendre contre les critiques. Mais le problème est désormais connu par nombre de philosophes analytiques, mais aussi par des économistes, des mathématiciens, des physiciens… Les papiers d'Elga et de Lewis sont

[6] Nick Wedd a sauvé ces discussions, encore disponibles sur la page http://www.maproom.co.uk/sb.html

scrutés à la loupe, les publications se multiplient. On se passionne, on se dispute, on ne se comprend plus devant cette énigme redoutable. L'énoncé qui suit, proche de celui d'Elga, est toutefois modifié pour ressembler aux énoncés les plus répandus, mais assurément aucune de ces modifications n'altère le problème :

La Belle au bois dormant :

Ce dimanche soir, des chercheurs vont endormir pendant quelques jours la Belle au bois dormant ; puis ils lanceront une pièce de monnaie équilibrée. Selon le résultat, ils interrompront brièvement le sommeil de la Belle soit une, soit deux fois : si face, un réveil le lundi ; si pile, un réveil le lundi, un autre le mardi. Chaque fois, ils auront un entretien avec elle, puis la rendormiront à l'aide d'une drogue qui lui fera complètement oublier ce réveil. Voici que la Belle, qui connait tout ce protocole, se réveille au cours de l'expérience, incapable de savoir si c'est lundi ou mardi. On lui demande alors : « À quel degré devez-vous croire que la pièce est tombée sur face ? » Que doit-elle répondre ?

Elga (2000) expose ainsi les deux réponses et les deux premiers raisonnements, intuitifs et imparfaits, qu'à la place de la Belle chacun d'entre nous pourrait accepter :

> *Première réponse* : 1/2, bien sûr ! Au départ, vous étiez certain que la pièce est équilibrée, donc le crédit que vous accordiez à l'issue face du tirage était 1/2. Après votre réveil, vous ne recevez aucune nouvelle information (vous saviez depuis le début que vous seriez réveillé). Donc le crédit que vous accordez à l'issue face devrait rester 1/2.
>
> *Seconde réponse* : 1/3, bien sûr ! Imaginez l'expérience répétée plusieurs fois. Alors, à long terme, environ un tiers des éveils seraient des *réveils-face* (des réveils qui se produisent lors d'expériences au cours desquelles la pièce tombe sur face). Donc, lors d'un réveil particulier, vous devriez croire au degré 1/3 que ce réveil est un réveil-face, et donc au degré 1/3 que la pièce de cette expérience

est tombée sur face. Cette considération reste en vigueur dans la circonstance présente, où l'expérience est effectuée une seule fois.

Le paradoxe est d'abord cette concurrence de deux arguments menant à des réponses incompatibles : on ne peut pas croire en face à la fois au degré 1/2 et au degré 1/3. Mais quand on regarde de près ce qu'écrit Elga, on s'aperçoit que la réponse 1/2, bien que tributaire d'une donnée physique (la pièce est équilibrée), est une probabilité épistémique déduite d'une autre probabilité épistémique grâce à une règle qui pourrait aisément prendre place dans une théorie de confirmation, ou plutôt ici de non-confirmation d'hypothèses : en absence d'information neuve, la probabilité *a posteriori* est égale à la probabilité *a priori*. En revanche, la réponse 1/3 est une probabilité ontique qui résulte d'un raisonnement typiquement fréquentiste : ce serait la fréquence (limite) des réveils-face parmi tous les réveils d'expériences répétées de nombreuses fois (ou une infinité de fois) ; ce n'est que parce que les expérimentateurs demandent un degré de croyance que la probabilité ontique se mue en crédit. Le paradoxe de la Belle au bois dormant serait par conséquent une rupture brutale du lien élastique entre probabilité épistémique et probabilité ontique, une scission qui va faire réagir : la plupart des analystes du problème vont en effet s'apercevoir, comme Elga l'a prévu, qu'il est facile d'être pro-1/2 quand on est bayésien, qu'il est facile d'être pro-1/3 quand on est fréquentiste, mais qu'il est difficile de réparer la scission ontique-épistémique, par exemple en élaborant un raisonnement bayésien favorable au 1/3 et en réfutant l'argument pro-1/2. L'inquiétude est telle qu'une centaine de publications en seulement quinze ans vont tenter de contribuer à la résolution du paradoxe dans une direction ou dans une autre. Il n'y a à ce jour aucun consensus.

Remarquons que, de la Belle au bois dormant, des chercheurs sont récemment revenus sur l'analyse plus ou moins abandonnée du Conducteur distrait : Walliser et Baratgin (2010) sont favorables au 1/3, tandis que

Schwarz (2015) et Gerville-Réache (2016) sont plutôt pro-1/2. Le Conducteur distrait profite des progrès accomplis dans l'étude de la Belle au bois dormant, qui apparait aujourd'hui comme une composante du premier problème sur laquelle les projecteurs sont braqués.

Il est clair que les deux positions temporelles de la Belle dans les mondes-pile sont les pièces importantes du puzzle. La Belle au bois dormant est un problème de *self-location*, c'est-à-dire un problème où un agent ou plusieurs agents ont besoin de *se* localiser dans l'espace-temps autant que dans l'espace logique, d'apprécier, bien souvent à l'aide des probabilités, où ils peuvent être, quand ils peuvent être, voire qui ils peuvent être. Tous les problèmes d'auto-localisation ne produisent pas un effet paradoxal. Néanmoins, nous pensons qu'il existe une dépendance insolite entre la tension ontique-épistémique et la perturbation de l'espace de probabilité d'un agent par sa prise en compte de propositions auto-localisantes ou, plus généralement, de propositions sensibles au passage du temps. La force d'un paradoxe peut être telle qu'elle relie confusément ce qui ne peut pas l'être *a priori*. À nous de ne pas être dupes, certes. Et pourtant, lorsque dans l'histoire est apparu un intérêt pour des possibilités, épistémiques ou ontiques, qui sortent de l'ordinaire, le ressort qui lie les deux interprétations a été très sollicité et usé. Cela se vérifie chez David Lewis, le plus connu des théoriciens de l'auto-localisation, comme, deux siècles avant lui, chez le sceptique de la théorie des probabilités Jean le Rond D'Alembert : lorsque le temps est devenu un paramètre essentiel des objets qu'ils probabilisaient, ces philosophes se sont préoccupés plus que jamais de la double interprétation, ils ont distingué deux probabilités pour un même objet, se sont demandé *quand* les deux valeurs devaient être égales ou pouvaient s'écarter l'une de l'autre. C'est assez extraordinaire chez D'Alembert parce que personne avant lui n'avait osé : les thèses balbutiantes de l'encyclopédiste bientôt fâché avec Diderot n'ont d'ailleurs

pas su convaincre[7]. Lewis, lui, semble arriver bien tard, alors que d'autres auteurs lui ont préparé le terrain ou l'ont devancé. Pourtant sa pensée est profonde, novatrice, assurée ; son principe d'alignement (on dit aussi de calibration ou de coordination des probabilités ontique et épistémique), appelé le « principe principal », n'est pas purement synchronique et sous-entend que l'alignement, à l'épreuve du temps, a une résistance limitée[8]. Subjectiviste convaincu, Lewis se serait-il attardé sur une probabilité ontique et objective s'il n'avait pas commencé à théoriser l'auto-localisation ? Ce n'est pas facile à dire mais, en tout cas, nous observerons comment le principe principal et d'autres règles d'alignement ou encore de révision des probabilités interviennent dans les débats autour des problèmes posés par les travaux du professeur de Princeton.

**Comment articuler double interprétation et auto-localisation ?**

Cinq chapitres seront nécessaires, même s'il est clair que la Belle au bois dormant est le point d'orgue et que les autres moments y conduisent lentement mais sûrement. Tous évoquent d'ailleurs le grand paradoxe de ce début de siècle d'une façon ou d'une autre.

Les deux premiers chapitres sont une plongée dans l'histoire et dans certains aspects ciblés des théories des probabilités, de la pluralité des mondes, de l'indexicalité du langage naturel, en somme de ce qui a permis qu'on invente à partir de la fin du 20$^{e}$ siècle des expériences de pensée où l'on doit estimer la probabilité de sa place dans l'univers, d'une date, de son identité. Il s'agit d'étayer la thèse de la dépendance des histoires et des théories et de leur convergence vers un paradoxe, mais pas de prouver

[7] Diderot (1761). Nous nous concentrerons non seulement sur le deuxième volume des *Opuscules mathématiques* de D'Alembert, mais aussi sur l'article « Croix ou pile » de l'*Encyclopédie*.

[8] Lewis (1980). Cf. Meacham (2010) et Bradley (2011b) pour des réflexions récentes sur le principe principal et le temps.

qu'auto-localisation et double interprétation seraient les deux faces d'une même préoccupation philosophique qui n'aurait jamais aperçu sa propre unité. Nous nous gardons bien de croire que théoriser deux genres de probabilités, *c'est* théoriser l'auto-localisation ; ce n'est pas juste un non-sens, c'est l'illusion extrême qu'il faut combattre, ce serait la pire leçon à tirer de l'étude d'un paradoxe. Le chapitre 1 se concentre sur des jalons d'une longue période, de Blaise Pascal au Carnap de la fin des années 1940, période durant laquelle, *a priori*, l'auto-localisation n'est presque rien, tandis que deux probabilités sortent peu à peu de leur confusion originelle ; il montre que D'Alembert déjoue les pronostics bien qu'il n'ait jamais songé à probabiliser sa position dans le temps. Le chapitre 2 est consacré à l'auto-localisation proprement dite et au bayésien épisodiquement propensionniste David Lewis, à ses références, en philosophie du langage autant qu'en logique et même en cosmologie, ainsi qu'aux théoriciens sur lesquels Lewis a eu une influence ; des exemples où des individus hésitent entre plusieurs localités ou plusieurs identités seront occasionnellement proposés.

Les trois derniers chapitres sont des analyses d'énigmes probabilistes où intervient spécifiquement la localisation *dans le temps*, et notamment de la Belle au bois dormant. Il s'agit d'observer comment réagissent aux difficultés l'approche bayésienne et l'approche fréquentiste, comment elles se rassurent l'une l'autre ou au contraire comment elles s'opposent. Le chapitre 3 prépare de plusieurs manières les analyses de *Sleeping Beauty* : résolution d'un problème à structure similaire en évitant le repérage dans le temps, généralisation des problèmes avec un dormeur amnésique, étude de plusieurs mini-problèmes de dormeur amnésique… Le long chapitre 4 présente le paradoxe de la Belle de manière à amener très naturellement les types de résolution que l'on trouve dans l'abondante littérature ; ceux-ci sont décryptés et commentés à tour de rôle, et l'on découvre que certains auteurs, parfois clairvoyants et astucieux, parfois effrontés et obscurs, ont à

leur façon cherché à réparer la scission ontique-épistémique désignée par Elga. Enfin, le chapitre 5, après analyse de plusieurs variantes et un passage par le raisonnement anthropique, défend une nouvelle solution qui n'est pas qu'une synthèse des vues précédentes, mais aussi une vraie pacification des deux interprétations de la probabilité. Le but est de clarifier, peut-être même de dénouer leur attache avec les réflexions sur le passage du temps et la localisation, mais aussi, peut-être, de tirer de l'aventure de la Belle une leçon touchant le Paradoxe suprême, toujours bien vivant, de la probabilité une et deux.

Juste une remarque à l'intérêt limité avant de commencer : les auteurs cités, leurs conventions, leurs notations sont tellement nombreux et variés que nous avons renoncé à un langage mathématique uniformisé dans toute la thèse, avec une lettre pour telle variable ou telle fonction et pas une autre. Nous pensons néanmoins être clair dans nos conventions et nos définitions. Apportons au moins cette bonne nouvelle : lorsque nous parlons d'une probabilité assurément ontique, elle est toujours notée par la lettre p *minuscule*. Les formes de la probabilité épistémique sont changeantes : c, C ou encore P. Les auteurs eux-mêmes font parfois des distinctions, comme celle du crédit initial C et du crédit actuel P. Il va de soi que la lettre C rappelle des mots comme confirmation, confiance, croyance ou crédit.

# Chapitre 1

# Deux probabilités

Plonger, même succinctement, dans l'histoire de la théorie des probabilités semble dispensable à une étude des interprétations de la probabilité dans la littérature sur l'auto-localisation, tout simplement parce que la croyance partielle auto-localisante est une idée très contemporaine, qui n'apparait que dans la seconde moitié du 20e siècle. Pourtant rien n'est simple lorsque nous remontons le temps ; des pistes ou des clés imprévues peuvent s'offrir à nous. Reconnaissons d'abord que Rudolf Carnap est loin d'être le premier auteur à différencier une probabilité épistémique et une probabilité ontique, même si cette différence, la différence radicale de *deux concepts*, est chez lui particulièrement soulignée. Au 19e siècle, Poisson et Cournot distinguaient deux probabilités ; au 18e siècle, Condorcet prenait déjà assez nettement cette direction[9]. Mais ce qu'a fait le maître de ce dernier, D'Alembert, est peut-être d'un plus grand intérêt pour la suite de cette thèse : en effet, ses réflexions sur les alignements et les écarts entre une probabilité dite physique et une probabilité dite métaphysique sont indissociables de ses tentatives de résolution de problèmes où il accordait au temps un rôle majeur, à commencer par ses analyses, incomprises à son

[9] Hacking (2002), p. 39. Le texte remonte là aussi le temps, de Carnap à Condorcet, avant de revenir soudain à Bertrand Russell.

époque, du jeu de « croix ou pile » qu'il était le seul à se représenter en paradoxe probabiliste. Nous pensons que D'Alembert avait pratiquement des préoccupations d'un théoricien de l'auto-localisation. Avant d'aborder ce sujet difficile qui doit dompter l'anachronisme et non y sombrer, il convient de se faire une idée du contexte philosophique de l'époque et de l'histoire de l'interprétation double de la probabilité, tant qu'à faire jusqu'à Carnap. Ce philosophe est moins déterminant que Frege dans la manifestation des théories de l'auto-localisation, bien qu'on retrouve une origine de la sémantique des mondes possibles dans son livre *Meaning and Necessity: A Study in Semantics and Modal Logic*. Mais surtout, la pensée de Carnap accompagne les découvreurs de célèbres principes d'alignement ontique-épistémique.

## 1. Brève histoire de l'interprétation double

Une préoccupation constante guide Ian Hacking dans *L'émergence de la probabilité* : montrer comment le 17e siècle, où ne se différencient ni plusieurs aspects ni plusieurs concepts de probabilité, va néanmoins lancer le « ressort » qui lie et délie, comme dans l'histoire de (dés)amour d'un couple tantôt déchiré et tantôt reconstruit, les interprétations de la probabilité. Il est plus juste, dans l'esprit de cette étude historique de Hacking qui, comme nous ici, ne se concentre sur le passé que pour comprendre des problèmes contemporains, d'affirmer que ce ne sont pas deux interprétations indépendantes que nous théorisons, comme si l'une avait pu naitre et perdurer sans l'autre, mais une interprétation historiquement double. Peut-être essentiellement double. Ce « peut-être » annonce la confrontation des théories de Poincaré et de Carnap, terme provisoire d'un survol historique qui commence avec une des figures chères à Hacking : Blaise Pascal.

### 1.1. Pascal, partis, pari

La thèse de l'interprétation historiquement double est quasiment vérifiée si l'on se tourne vers le premier grand penseur de la probabilité. C'est en tant que mathématicien que Pascal correspond avec Fermat en 1654 et expose ce « problème des partis » qui hante les joueurs depuis longtemps : comment deux joueurs qui interrompent leur jeu en trois (ou quatre, ou cinq…) parties gagnantes sans qu'aucun des deux n'ait remporté trois (ou quatre, ou cinq…) victoires doivent-ils partager les mises engagées au début ? Dans sa réponse, Pascal considère les événements qui auraient pu se produire avec un « hasard égal » si le jeu s'était poursuivi, et pour chacun d'eux les gains de chaque joueur, afin de calculer un gain moyen : il jette donc les bases du calcul des « chances » en associant gains et hasard. Il décède huit ans plus tard. Si l'histoire s'arrêtait là, nous le présenterions sans hésiter comme un objectiviste peu aguerri et pourtant mû par cette étincelle géniale que l'on a si souvent perçue en lui.

Mais il y a mieux : en 1670 paraissent les *Pensées*, où l'on trouve le passage souvent appelé « le pari de Pascal »[10]. En apologiste original, le philosophe invite les lecteurs qui doutent de l'existence de Dieu, plus exactement qui n'ont pas assez de raisons de croire en Dieu ni assez de raisons de ne pas croire en lui, à prendre une décision en pesant les conséquences : ils peuvent parier sur son existence ou bien parier sur son inexistence, de la même façon qu'ils parieraient sur pile ou sur face (« croix ») :

> Dieu est ou il n'est pas ; mais de quel côté pencherons-nous ? la raison n'y peut rien déterminer. Il y a un chaos infini qui nous sépare. Il se joue un jeu à l'extrémité de cette distance infinie, où il arrivera croix ou pile. Que gagerez-vous ? par raison vous ne pouvez faire ni l'un ni l'autre, par raison vous ne pouvez défaire nul des deux.

[10] Numéro 418 dans la numérotation de Lafuma, 233 dans celle de Brunschvicg.

Quand l'insuffisance des raisons de croire et de ne pas croire mène à l'indifférence, il faut considérer les gains et les pertes possibles. À leur mort, si les « joueurs » ont cru en Dieu, deux éventualités se présenteront : Dieu est, alors ils iront au paradis et ainsi auront un gain infini ; Dieu n'est pas, alors ils retourneront au néant en ne perdant rien. Mais s'ils n'ont pas cru en lui, les éventualités sont modifiées : si Dieu est, ils retourneront au néant ou, pire, iront en enfer et perdront tout ; si Dieu n'est pas, ils retourneront au néant en ne perdant rien. Le choix qui s'impose est donc le pari sur l'existence de Dieu, puisqu'il n'y a rien à perdre et tout à gagner.

Ce qui est pour nous important dans cet argument pascalien, ce n'est pas sa justesse, c'est le débordement de la théorie des jeux naissante : elle devient théorie de la décision et un de ses raisonnements est appliqué à la croyance en l'existence de Dieu, premier exemple (et audacieux exemple) de conjecture « probabilisable » qui ne saurait être vérifiée par l'expérience. La voie est ouverte vers d'autres inférences probabilistes de même genre. Hacking est de cet avis :

> Ces fragments […] montraient comment l'arithmétique de l'aléatoire pouvait faire partie d'un « art de conjecturer » plus général. Grâce à eux, il devint possible de comprendre que la structure d'un raisonnement sur le hasard pouvait être transférée à une inférence ne reposant pas sur un processus aléatoire.[11]

Le chercheur rappelle que Pascal n'emploie jamais les termes « probabilité » et « degré ». Le pari sur Dieu n'a pourtant qu'une explication :

> Ce qu'il dit, c'est que l'on se trouve dans la même attitude *épistémique* qu'un joueur face à une pièce dont les propriétés aléatoires sont inconnues. Le jugement de Pascal repose sur la supposition d'un isomorphisme entre la structure d'un problème décisionnel où l'on connait l'existence de chances

[11] Hacking (2002), p. 101.

> physiques objectives et celle d'un problème où il n'y a pas de chances physiques objectives.[12]

Ainsi sont reliés le plan physique, où l'on devrait être capable de mesurer les probabilités des deux issues possibles du lancer d'une pièce de monnaie ou d'un autre instrument aléatoire (0,53 pour pile et 0,47 pour face par exemple), et le plan psychique où l'on devrait être capable de croire en une hypothèse ou en son contraire… à un certain degré, si le raisonnement était parachevé. On voit que, même si Pascal ne demande pas d'évaluer quelque chose comme un degré de croyance en pile ou en face (dans l'ignorance des propriétés de la pièce), ou un degré de croyance en Dieu (devant le non-sens des « chances que Dieu existe »), l'idée est prête à émerger, à sortir sous sa plume. Il ne manque qu'un pas que la *Logique* de Port-Royal, marquée notamment par la pensée pascalienne comme chacun sait, franchit en représentant sur une échelle numérique la probabilité épistémique et en utilisant par ailleurs l'expression « degré de probabilité » qui se popularisera[13].

Pascal n'a donc pas eu le temps d'être un objectiviste radical : le passage au subjectivisme par déplacement impérieux de structures de problèmes d'un plan à l'autre fut immédiat. Cependant, nous ne saurons jamais à quel point il fut harmonieux et naturel. Nous pouvons par contre observer que les prémisses et les conclusions de l'argument apologétique auraient pu compromettre (mais n'ont pas compromis) la réception critique de cette probabilité venue du monde sensible et accomplie hors de lui : alors qu'on commenta souvent élogieusement et qu'on porta un immense intérêt à la solution du problème des partis, le pari sur l'existence de Dieu devint un texte contre lequel il fallait prendre position. Certes, quelques religieux le trouvaient astucieux ; Locke fut une des rares autorités à

[12] *Ibid.*, p. 109.

[13] *Ibid.*, chap. 9.

l'apprécier[14]. Mais il faut reconnaitre que cette manière de faire des terrains ontologique et épistémique un seul « terrain de jeu » n'a presque jamais convaincu, ni les joueurs concernés, ni les logiciens, ni les théologiens, ni les libres penseurs. Voltaire jugeait en outre inconvenant de décider de l'existence d'un Être suprême en fonction de considérations d'intérêt.

Le pari de Pascal présente un raisonnement valide au sens logique (c'est sa force) qui cherche à prouver avec une certaine décontraction tout le bénéfice que peut apporter la foi en Dieu, grâce à une analogie avec un simple jeu de hasard, ce qui rebute notre esprit, mais pas au point de provoquer un effet paradoxal durable. Peu de personnes, aujourd'hui comme hier, sont persuadées ou déconcertées par l'argument. L'effacement des difficultés soumises par Pascal à l'esprit critique de ses lecteurs fut quasiment immédiat, car une avalanche d'objections presque toujours soignées et unanimement reçues montra l'imprudence du philosophe qui, parti en éclaireur sur le chemin encore mystérieux d'une théorie des probabilités, construisit un raisonnement valide sur des prémisses néanmoins inacceptables, l'imprégna de sa culture chrétienne particulière, anticipa une vie dans l'au-delà en oubliant celle d'ici-bas… Des présupposés non universels et des choix discutables modifient la conclusion finale ; dès que quelques-uns sont identifiés, la réfutation de l'argument devient facile. Mais il faut remarquer qu'en lui-même le passage d'une probabilité d'événements du monde à une probabilité de conjectures expérimentalement invérifiables n'est jamais contesté, peut-être parce que personne n'en prend suffisamment conscience. Aussi, l'imprudence de Pascal n'a pas freiné le futur développement de théories sur une probabilité à plusieurs visages.

[14] *Ibid.*, p. 109.

### 1.2. Du 18e au 20e siècle

Durant le siècle des Lumières, les philosophes entremêlent encore dans leurs écrits les approches ontologique et épistémologique. Cependant, Bayes et Laplace commencent à se démarquer, et en réaction au subjectivisme qui attire indéniablement ces autorités apparait, principalement chez les empiristes anglais, un discours opposé soulignant de plus en plus nettement qu'une probabilité doit être lue dans le monde, expérimentalement vérifiée. Condorcet est peut-être le moderne qui prend le plus conscience de ce double visage de la probabilité. Il appelle « facilité » la face ontique et « motif de croire » la face épistémique. Séduit par le sensualisme d'outre-Manche, il articule les deux faces d'une manière originale et subtile bien expliquée par Bernard Bru, qui n'hésite pas à parler d'une « école fréquentiste subjectiviste » :

> Supposons que d'une urne contenant des boules blanches et noires en proportion inconnue, on ait tiré chaque matin depuis une éternité, mettons 99 999 fois, une boule blanche, on serait assuré, avec une certitude presque aussi forte que la précédente, que lorsque demain on tirera une boule de l'urne, elle sera nécessairement blanche. Or la théorie de l'estimation de Condorcet nous apprend que l'urne la plus raisonnable que l'on puisse déduire de ces 99 999 observations est composée comme tout à l'heure de 100 000 boules blanches et 1 boule noire. Évidemment il ne s'agit pas de l'urne absolument vraie, celle-là est hors de portée, mais d'une « urne moyenne », qui possède le plus de propriétés statistiques et de vertus probabilistes que l'on puisse espérer à court terme comme sur le très long terme […].
>
> C'est donc que d'une telle urne moyenne […] on a le même motif de croire que l'on va tirer une boule blanche que celui que l'on déduirait de la permanence de ces 99 999 observations identiques, motif qui se trouve ainsi mesuré par la probabilité calculée de la façon qu'on a dite. Et notre motif de croire augmentera avec la probabilité calculée comme il augmente avec le nombre d'observations identiques d'un même phénomène, avec la plus grande permanence de nos sensations. Ainsi puisque toutes nos certitudes dérivent de la permanence de nos

> sensations, elles peuvent être mesurées par un calcul de probabilité qui devient véritablement une mesure de nos motifs de croire. […]
>
> Quant à ces « combinaisons également possibles », elles-mêmes ne sont réelles que par l'observation répétée que l'on pourrait faire qu'elles apparaitraient également sur le long terme, l'urne réelle, dont on les tirerait, n'étant, en réalité, qu'une fiction mathématique particulièrement propice aux calculs et à l'intuition (et aux illusions) du géomètre. On peut ainsi considérer que Condorcet est le fondateur de l'École « fréquentiste subjectiviste » […].[15]

Bru estime que l'« idéalisme statistique » de Condorcet est une pensée cohérente qui affaiblit le préjugé selon lequel la philosophie des Lumières mélangerait de manière contradictoire le rationalisme cartésien et l'empirisme lockien. Notre accord avec ce commentaire de Bru est bien réel, malgré ici notre récit éclair de deux siècles d'histoire.

Au 19e siècle, quelques mathématiciens français, dont Cournot, témoignent des difficultés et des malentendus rencontrés par les chercheurs et soulignent un double sens du mot « probabilité ». Cela peut paraitre étonnant chez Cournot, qui est longtemps resté dans les mémoires en tant qu'objectiviste et parfois fréquentiste ; pourtant une telle réduction (voire déformation) de sa théorie n'est plus d'actualité, principalement parce que la place qu'il fait à la probabilité épistémique est reconnue[16]. C'est aussi lui qui se risque à la plus claire des distinctions entre la « chance », mesure objective de la possibilité d'un événement, qui appartiendrait plutôt à une théorie de la nature, et une probabilité subjective, relative « en partie à nos connaissances, en partie à notre ignorance », qui relève donc d'une théorie de la connaissance[17]. Malgré tout, cette distinction est peut-être davantage

[15] Bru (1994), p. 11-12.

[16] On comparera Bru (1994) et Martin (1994) pour entrevoir ce qu'il en est vraiment de ce « fréquentisme » d'une part, de la dette de Cournot envers Condorcet d'autre part. Martin (1996) est plus complet sur ces sujets.

[17] Cournot (1843), p. 155 : « de nombreuses équivoques […] se rectifient dès qu'on a présente à l'esprit la distinction fondamentale entre les probabilités qui ont une

l'éclosion de deux fleurs sur la même branche qu'une rupture conceptuelle radicale, chacune des deux notions ayant besoin de sa sœur pour être bien entendue, voire bien quantifiée.

La plupart des philosophes du 20e siècle qui perçoivent l'ambiguïté probabiliste s'efforcent de la faire disparaitre en montrant, ou bien qu'un concept unique a un visage double, ou bien que deux concepts passent à tort pour un seul. Certains théoriciens radicaux, comme le fréquentiste Neyman ou le bayésien classique de Finetti, pensent que seule une fréquence d'événements aléatoires ou au contraire seul un degré de croyance devrait être appelé « probabilité » : la question de l'alignement entre l'ontique et l'épistémique n'a aucun sens pour eux. En revanche, la question peut être enfin posée aux philosophes ouverts aux deux interprétations, qu'ils aient ou pas tendance à en privilégier une ; malheureusement, nous allons constater que les réponses apportées sont souvent effrontément simples[18].

### 1.3. Poincaré et le pari de la double interprétation

Poincaré est une source essentielle pour beaucoup de penseurs de la probabilité au 20e siècle. Le fréquentiste Reichenbach et le bayésien de Finetti ont fait de ses textes des lectures si orientées qu'elles s'opposent diamétralement[19]. Ce fait remarquable est sans doute la conséquence de la manière dont Poincaré expose sa pensée. Dans son esprit, il n'y a qu'un concept de probabilité, mais il utilise ce concept parfois plutôt dans un

existence objective, qui donnent la mesure de la possibilité des choses, et les probabilités subjectives, relatives en partie à nos connaissances, en partie à notre ignorance, variables d'une intelligence à une autre, selon leurs capacités et les données qui leur sont fournies. »

[18] C'est à Barberousse (2000) que nous empruntons l'opposition des deux figures Poincaré-Carnap.

[19] Bienvenu (2007), p. 89-92 ; Gillies (2000), p. 85-87.

sens, parfois plutôt dans un autre, en s'adaptant aux situations. Il est convaincu qu'une claire distinction de deux interprétations est notamment nécessaire pour apporter des solutions aux paradoxes et aux cercles vicieux que nous rencontrons en définissant la notion. Dans *La Science et l'Hypothèse*, il résume ainsi ces difficultés majeures :

> Le nom seul de calcul des probabilités est un paradoxe : la probabilité opposée à la certitude, c'est ce qu'on ne sait pas, et comment peut-on calculer ce que l'on ne connaît pas ? Cependant, beaucoup de savants éminents se sont occupés de ce calcul, et l'on ne saurait nier que la science n'en ait tiré quelque profit. Comment expliquer cette apparente contradiction ?
>
> La probabilité a-t-elle été définie ? Peut-elle même être définie ? Et, si elle ne peut l'être, comment ose-t-on en raisonner ? La définition, dira-t-on, est bien simple : la probabilité d'un événement est le rapport du nombre de cas favorables à cet événement au nombre total des cas possibles. […] Pourquoi la première manière d'énumérer les cas possibles est-elle plus légitime que la seconde ? En tout cas, ce n'est pas notre définition qui nous l'apprend. On est donc réduit à compléter cette définition en disant : « … au nombre total des cas possibles, pourvu que ces cas soient également probables ». Nous voilà donc réduits à définir le probable par le probable.[20]

La probabilité est d'abord, en quelque sorte, une dose de certitude dans une croyance qui n'est que candidate au rang de connaissance. Le calcul d'une probabilité parait suspect parce qu'il apporte de la certitude à de l'incertitude, parce qu'il est détermination objective de ce qui semble d'abord être estimation subjective. En outre, calculer le rapport du nombre des cas favorables sur le nombre total des cas ne fait pas disparaitre la dimension subjective, comme le montre le problème de l'équiprobabilité. À l'inverse, la dimension objective parait elle aussi nécessaire, et tout agent qui prend ses décisions en pariant sur leur succès est confronté un jour ou l'autre au « phénomène » de la conformité des règles du calcul avec la répartition des événements dans une longue série :

[20] Poincaré (2013), p. 131-132.

> Je vois bien ce qu'on pourrait dire : « Nous sommes ignorants et pourtant nous devons agir. Pour agir, nous n'avons pas le temps de nous livrer à une enquête suffisante […]. Nous devons donc nous décider sans savoir ; il faut bien le faire au petit bonheur et suivre des règles sans trop y croire. Ce que je sais, ce n'est pas que telle chose est vraie, mais que le mieux pour moi est encore d'agir comme si elle était vraie ». Le calcul des probabilités, et par conséquent la science, n'aurait plus qu'une valeur pratique.
>
> Malheureusement la difficulté ne disparaît pas ainsi : un joueur veut tenter un coup ; il me demande conseil. Si je le lui donne, je m'inspirerai du calcul des probabilités, mais je ne lui garantirai pas le succès. C'est là ce que j'appellerai la *probabilité subjective*. Dans ce cas, on pourrait se contenter de l'explication que je viens d'esquisser. Mais je suppose qu'un observateur assiste au jeu, qu'il en note tous les coups et que le jeu se prolonge longtemps ; quand il fera le relevé de son carnet, il constatera que les événements se sont répartis conformément aux lois du calcul des probabilités. C'est là ce que j'appellerai la *probabilité objective* […].[21]

La probabilité subjective est indissociable de la probabilité objective : une seule suite d'inférences et un seul calcul tendent à objectiver notre disposition subjective à parier, qui finit par égaler les proportions et les fréquences des événements du monde sous certaines conditions. Le monde fournit même des informations qui affinent des probabilités dont nous ne pouvons pas nous rendre maîtres par le calcul. Nous avons bien affaire, chez Poincaré, à un monisme du concept mais à un dualisme de l'interprétation. Quand nous affirmons que la probabilité d'obtenir 6 avec un dé est 1/6, nous voulons dire que cette valeur mesure une propriété observable, que le 6 apparaitrait vraisemblablement une fois sur six lancers consécutifs, et nous voulons aussi dire que, lors d'un lancer particulier, nous croyons peu à la venue d'un 6, nous sommes prêts à parier cinq fois moins d'argent sur 6 que sur non-6 si nous fondons ce pari sur la probabilité objective. Parfois, l'interprétation subjectiviste permet de mieux

[21] *Ibid.*, p. 134.

justifier une estimation probabiliste ; dans d'autres contextes, l'interprétation objectiviste est préférable ; pourtant, tous les énoncés de probabilité entrent dans un seul discours scientifique. C'est parce qu'une interprétation peut ainsi dominer l'autre dans l'exposition de sa pensée que Poincaré est parfois considéré comme un des premiers pluralistes contextuels[22].

Un partisan de la probabilité à deux visages comme Poincaré peut tout au plus excuser un écart entre un degré de croyance et une fréquence (ou une propension) mais en aucun cas le *justifier*. Il peut arriver que la probabilité épistémique ne puisse être mise en relation avec la probabilité ontique que de manière très lâche, parce que les calculs sont compliqués ou parce qu'on n'a pas pu envisager tous les cas possibles. Mais, idéalement, probabilités épistémique et ontique sont parfaitement alignées ; il est impossible de formuler la loi d'une divergence. Comment pourrait-il en être autrement ? La probabilité n'est degré de croyance ou n'est fréquence qu'interprétativement. Puisque le non-alignement est accidentel et non rationnel, la question de l'alignement est vite résolue.

### 1.4. Carnap et le pari du double concept

La philosophie analytique du milieu du 20e siècle comprend mal le monisme-dualisme de Poincaré pour une raison évidente : le discours sur la probabilité à deux visages demande aux esprits rigoureux de ne penser qu'un concept mais son admission logique les oblige à détacher malgré tout deux concepts. Il nous semble que le lecteur dubitatif, confronté à une probabilité étrange qui apparait tantôt comme une propriété des croyances ou des hypothèses, tantôt comme une propriété d'objets physiques, se retrouve face à deux choix, deux moyens de clarifier enfin la situation : ou bien proposer une métaphysique subtile solidarisant ou réunissant d'une

[22] Par exemple par Bienvenu (2007), p. 12.

façon ou d'une autre l'esprit et la matière, un monde des croyances et un monde des faits objectifs ; ou bien abandonner l'unicité du concept. La seconde voie est celle des logicistes hostiles à la métaphysique. Carnap est de ceux-là.

L'illustre membre du Cercle de Vienne a rédigé des textes de référence sur l'ambiguïté probabiliste, notamment plusieurs articles dont *The Two Concepts of Probability* publié en 1945, mais aussi *Logical Foundations of Probability*, un ouvrage volumineux paru en 1950 qui constitue un traité de logique inductive encore incomplet. Dans ces textes, un concept épistémique et un concept ontique, irréductibles l'un à l'autre, sont expliqués : la probabilité logique, qui est le degré de confirmation d'une hypothèse par un ensemble de données et qui est simplement appelée probabilité$_1$, est distinguée de la probabilité statistique ou probabilité$_2$, associée à une fréquence relative d'événements d'un certain type dans une longue série d'événements. Il convient de préciser que cette conception carnapienne d'une probabilité épistémique et néanmoins objective (probabilité$_1$), qui se situe dans la lignée de John Maynard Keynes, se transformera beaucoup après vingt années de recherches intensives pour finalement se rapprocher de la conception subjectiviste, personnaliste de Leonard Savage[23]. Nous nous concentrons ici exclusivement sur la première théorie de Carnap.

Pour le professeur à l'université de Chicago, les malentendus sur la probabilité commencent dans le langage naturel. En anglais comme en français, en latin comme en allemand, des mots à double sens, à commencer par « probable », dissimulent la dualité de la notion. Un mot comme « vraisemblable », qui pourrait n'être employé que pour qualifier

[23] Paquette (2002), p. 66-68 ; Carnap (2015), introd. de P. Wagner, p. 11-14, 40. Pierre Wagner précise que, dans les derniers écrits sur la logique inductive, l'orientation objectiviste intéresse toujours Carnap, qui n'envisage que des degrés de croyance rationnels.

des hypothèses ou d'autres propositions logiques, tend lui aussi à qualifier les faits ou les états décrits et à prendre tous les sens de « probable ». C'est pourquoi quelques auteurs ont été mal compris quand ils ont voulu mettre à profit cette disponibilité de plusieurs mots dans le langage ordinaire pour nommer les concepts qu'ils distinguaient. Les théories de la probabilité se sont développées dans de mauvaises conditions, accompagnées de nombreuses querelles. Cependant, il est possible aujourd'hui de discerner dans la littérature une rivalité principale entre deux groupes de penseurs : ceux qui considèrent une certaine relation logique entre les propositions, et ceux qui se concentrent sur une fréquence relative d'événements[24]. Les choses sont claires pour Carnap : il y a deux concepts différents, qui appartiennent à des théories différentes mais pas incompatibles. C'est en quelque sorte une bonne nouvelle, l'annonce de la fin des hostilités ; les philosophes qui ont discerné le dualisme du concept peuvent désormais travailler tranquillement et indépendamment à une théorie logique ou bien à une théorie statistique.

Carnap distingue donc probabilité$_1$ et probabilité$_2$, il utilise les indices 1 et 2 pour rappeler que, malgré l'homonymie, résidu nécessaire du langage naturel, la distinction est radicale. Apportons encore des précisions. « Le jet d'un dé » est une classe d'événements ; « le jet de ce dé à minuit » est un événement. Quelques auteurs appellent « événement » ce qui est en réalité une classe d'événements, et ainsi entretiennent la confusion des deux concepts de probabilité. En effet, si l'on considère que probabilité$_1$ et probabilité$_2$ sont des fonctions de deux arguments, ceux-ci sont bien différents selon le concept envisagé. Les arguments de probabilité$_1$ sont une

[24] Carnap reconnait que, dans le premier groupe, se rangent aussi des philosophes qui soulignent la subjectivité de certaines estimations probabilistes, mais il pense que des relations logiques ne peuvent être qu'objectives. En revanche, il ne dit pas un mot des avocats de la probabilité-propension. La raison évidente, dirons-nous, est que cette approche ne fait son apparition que plus tard, grâce à Popper ; pourtant elle est préfigurée dans des écrits de Peirce, un auteur cité par Carnap à d'autres occasions.

hypothèse et un ensemble d'informations, autrement dit les expressions de faits ou d'événements ; en revanche, probabilité$_2$ se réfère à deux classes d'événements, elle donne la fréquence des événements de la première classe, ayant une certaine propriété, parmi les événements réunis par une autre propriété dans la seconde classe. Il s'agit là d'une différence fondamentale entre les deux concepts. Une autre différence importante est que les énoncés qui contiennent le concept de probabilité$_1$ sont tous analytiques, le degré de confirmation d'une hypothèse par une information étant déterminé par la seule analyse logique des propositions et de leurs relations, alors que les énoncés de probabilité$_2$ les plus simples, qui sont à distinguer des axiomes et des théorèmes de la théorie mathématique de probabilité$_2$, sont empiriques, leur vérification passant nécessairement par l'observation, dans la nature, de faits pertinents (les résultats d'un grand nombre de lancers de pièces ou de dés par exemple).

Le concept purement logique de probabilité$_1$ est une pièce maîtresse pour le projet carnapien d'élaboration d'une logique inductive qui aurait les mêmes pouvoirs que la logique classique déductive, qui serait une branche à part entière de la logique. Le projet s'est soldé par un échec, ce qui est très loin de rendre inintéressante l'analyse du concept et de son homonyme probabilité$_2$. Simplement, on peut comprendre pourquoi Carnap insiste sur l'indépendance des énoncés de probabilité$_1$ vis-à-vis de la « contingence des faits »[25] et donc des fréquences d'événements : il veut que sa logique inductive repose sur un calcul propositionnel qui ne se soucie jamais de la réalité des faits décrits. Nous sommes naturellement conduits à supposer que, dans ces conditions de séparation stricte entre l'épistémique et l'ontique, jamais aucune probabilité d'un des deux genres ne peut être alignée dans une inférence rigoureuse sur une probabilité de l'autre genre ;

[25] Cf. Carnap (2015), p. 58 : « Un *énoncé* simple de probabilité$_1$ […] est indépendant de la contingence des faits parce qu'il ne dit rien sur les faits (bien que les deux arguments réfèrent bien, en général, à des faits). »

nous supposons donc que, sur ce point précis, nous avons trouvé en Carnap un anti-Poincaré, un philosophe pour qui, certes, la question de l'alignement a un sens, qui y répond, certes, en une sentence lapidaire, mais une sentence diamétralement opposée à celle du savant français. Avons-nous raison ? Pas exactement.

### 1.5. L'alignement ontique-épistémique chez Carnap

Il faut reconnaitre que Carnap prend au sérieux la question de l'alignement, qu'il connait principalement grâce à ses lectures de Reichenbach ; sa réponse est nuancée et il ne fait pas comme si elle allait de soi. Le long chapitre IV des *Logical Foundations* s'applique à montrer les liens que l'on peut tisser entre les deux concepts de probabilité, et notamment la tentation d'expliquer probabilité$_1$ comme l'*estimation* d'une fréquence relative, ou la bonne entente des deux concepts dans des argumentations où ils sont impliqués. L'article *The Two Concepts of Probability* n'est pas en reste. Au détour d'une discussion sur le principe de vérifiabilité, cher aux positivistes logiques, et sur l'analyse selon laquelle « l'usage d'énoncés de probabilité$_1$ ne peut pas en soi violer le principe de l'empirisme »[26], Carnap condamne tout raisonnement qui voudrait conclure une fréquence par alignement sur une probabilité purement logique, en arguant que cette conclusion est factuelle : un énoncé factuel de probabilité$_2$ ne devient pas logique parce qu'un raisonnement l'aurait conclu, un énoncé de probabilité$_1$ ne devient pas factuel parce qu'il permettrait d'inférer une fréquence ; en réalité, le raisonnement mélange par inadvertance deux sortes de probabilités et doit être considéré comme aberrant. La transition inverse, de probabilité$_2$ vers probabilité$_1$, pose quelques soucis que Carnap propose de surmonter par l'étude d'une situation simple :

[26] Carnap (2015), p. 63.

> Je reconnais que dans certains cas, il existe une relation étroite entre la probabilité$_1$ et la fréquence relative. Néanmoins, la question décisive concerne la nature de ce lien. Considérons un exemple simple. Supposons que la donnée *e* dise que parmi 30 choses observées ayant la propriété $M_1$, on en a relevé 20 ayant la propriété $M_2$, de sorte que la fréquence relative de $M_2$ par rapport à $M_1$ dans l'échantillon observé est 2/3 ; supposons en outre que *e* dise qu'un individu particulier *b*, n'appartenant pas à l'échantillon, a la propriété $M_1$. Soit *h* la prédiction selon laquelle *b* possède $M_2$. Si le degré de confirmation *c* est défini d'une manière appropriée comme *explicatum* pour la probabilité$_1$, *c*(*h*, *e*) sera égal ou proche de 2/3. Supposons, pour simplifier, que *c* = 2/3. Pour autant, le fait que, dans ce cas, la valeur de *c*, ou probabilité$_1$, soit égale à une certaine fréquence relative n'implique nullement que la probabilité$_1$ soit ici la même chose que la probabilité$_2$. Ces deux concepts demeurent fondamentalement différents, même dans ce cas.[27]

Notons déjà que la fréquence relative 2/3 calculée après observation d'un échantillon très limité n'est pas une probabilité$_2$ selon la définition exacte donnée par Carnap, elle n'est pas la fréquence relative des objets ou événements ayant une propriété $M_2$ parmi ceux qui ont une propriété $M_1$ *dans l'entière séquence des événements pertinents*. Cette dernière fréquence est inconnue. Mais tout de même, la fréquence ou, si on préfère l'appeler ainsi, la proportion 2/3 est bien une propriété ontique ou physique, elle est assignée à deux groupes d'objets trouvés dans un échantillon, un extrait de la nature. Carnap a parfaitement conscience de tout cela et commence par admettre que, dans l'énoncé de probabilité$_1$ « *c*(*h*, *e*) = 2/3 », la valeur 2/3 est « d'une manière ou d'une autre basée sur notre connaissance empirique »[28]. Pourtant, seule la proposition *e* est factuelle : « *c*(*h*, *e*) = 2/3 » n'étant pas une répétition de *e* mais l'énoncé d'une relation logique entre *e* et *h*, il est incorrect de prétendre que celui-ci

[27] *Ibid.*, p. 64.

[28] *Ibid.*, p. 64.

doit être interprété comme énonçant la fréquence relative et comme étant lui-même factuel. Un passage, cependant, jettera peut-être le trouble :

> Comme on l'a dit, une estimation de la probabilité$_2$, la fréquence relative dans la suite entière, est basée sur la fréquence relative observée sur l'échantillon. Je pense qu'en un sens, l'énoncé « $c(h, e) = 2/3$ » peut lui-même être interprété comme exposant une telle estimation ; il dit la même chose que : « La meilleure estimation de la probabilité$_2$ de $M_2$ par rapport à $M_1$ selon la donnée $e$ est 2/3 ». Si d'aucun voulait appeler ceci une interprétation de la probabilité$_1$ en termes de fréquence, je n'y verrais pas d'objection.[29]

Carnap ne contredit pas sa précédente argumentation, il n'écrit pas que la proportion 2/3 des $M_2$ dans les $M_1$, dont on a rappelé le caractère factuel, approche la probabilité$_2$ des $M_2$ parmi les $M_1$, ce qui voudrait dire qu'une probabilité$_1$ identifiée avec une telle estimation serait elle-même factuelle ; il assure en revanche que sur la base de la proportion 2/3 on peut estimer la probabilité$_2$, qu'une formulation de la meilleure estimation suivant la donnée $e$ est purement logique, et qu'un énoncé de probabilité$_1$ peut être regardé comme une telle formulation. Néanmoins, le tour de force de Carnap, qui consiste à remettre la probabilité$_2$ au sein de la reformulation d'un énoncé de probabilité$_1$ afin de montrer la connexion des deux concepts pourtant toujours fondamentalement différents, semble avoir un prix qui est la mise à l'écart d'un des arguments de $c$, $h$. Celui-ci rend possible la reformulation, mais au final seul l'argument $e$ est apparent. Cela nourrit le sentiment d'une réinterprétation profonde du concept de probabilité$_1$ qui, en tant qu'estimation et non plus degré de confirmation, serait comme penché à l'extrême sur les fréquences relatives sans jamais s'identifier à elles.

En principe, la bipartition analytique/synthétique, un des « dogmes de l'empirisme logique » critiqué par Quine, n'autorise pas une inférence

[29] *Ibid.*, p. 65.

directe entre probabilité statistique et probabilité logique. Mais nous venons de voir que Carnap trouve une parade en renouant pour l'occasion avec la notion d'interprétation : reconnaitre que la probabilité$_1$ peut être fondée et même alignée sur une fréquence relative, c'est reconnaitre que le concept, bien qu'indivisible, est complexe, qu'une interprétation fréquentielle vient le diversifier, si ce n'est le rendre ambigu. Le positiviste donne au problème de l'alignement une solution incomplète temporaire, en pensant que, dans un futur proche, toute éventuelle ambiguïté doit disparaitre dans l'explication achevée du concept et la présentation des lois de la logique inductive.

### 1.6. À la recherche du philosophe moderne dissident

Notre rappel historique montre une évolution lente des conceptions de la probabilité. Une interprétation double de la probabilité pascalienne peut être faite aujourd'hui dans nos lectures du savant clermontois. Pourtant, il faut attendre le 19$^{e}$ siècle pour que les deux faces de la probabilité soient explicitement dissociées dans un traité sur le sujet. En outre, même au début du 20$^{e}$ siècle, les philosophes les plus conscients de l'ambiguïté probabiliste se préoccupent peu de la question émergente de l'alignement entre probabilités ontique et épistémique. Bien entendu, durant tout ce temps, l'auto-localisation n'est pas un problème pour la philosophie. Somme toute, notre rappel historique est assez convenu. Maintenant, ne pouvons-nous pas trouver, quelque part dans l'histoire des théories de la probabilité, un événement qui défie notre chronologie bien réglée, tel que la distinction prématurée, chez un auteur du 17$^{e}$ ou du 18$^{e}$ siècle, de deux genres de probabilités, accompagnée par une réflexion présentant des points communs, peut-être peu évidents de prime abord, avec les analyses très contemporaines de problèmes d'auto-localisation ? Nous n'osons pas croire à la possibilité d'un tel événement. Quand aurait-il eu lieu ? Qui l'aurait produit ?

Notre première envie est de relire d'illustres théoriciens qui ont durablement marqué les esprits et sont appréciés autant par les objectivistes que par les subjectivistes. Jacques Bernoulli n'est pas le moindre. Son *Ars Conjectandi* a éclairé tout le 18e siècle et continue de nos jours d'étonner et d'inspirer ; son approche de ce que l'on appellera bientôt la probabilité inverse, ou probabilité des causes par les événements, annonce les théorèmes de Bayes et de Laplace ; sa formulation de la loi des grands nombres prépare les futures théories fréquentistes[30]. Pourtant nous ne trouverons pas ici l'événement que nous cherchons. En réalité, Bernoulli est un de ces génies reconnus qui écrit l'histoire : il n'est pas en avance sur elle au point d'être à côté d'elle et de se différencier sur certains points. Il a existé en revanche des hommes qui auraient pu écrire une histoire parallèle mais dont certaines idées ont été longtemps détestées et enterrées parce qu'elles ne convenaient pas à la marche de l'histoire des idées, nécessairement unique en ce monde. Il en est ainsi d'Eubulide de Milet, cet adversaire d'Aristote inventeur de paradoxes philosophiques toujours non résolus aujourd'hui, qui aurait dû naitre à la fin du 19e siècle pour se sentir un peu mieux chez lui[31]. Il y a aussi des moutons noirs dans l'univers impitoyable des probabilités. Jean le Rond D'Alembert en fait partie ; reste à savoir s'il détient ce que nous voulons.

[30] Cf. Hacking (2002), p. 199 : « On en a fait le père de la première conception subjectiviste de la probabilité, ce qui n'a pas empêché Richard von Mises d'en faire un fréquentiste exemplaire. Des statisticiens plus récents, tels A. P. Dempster, disent de lui qu'il anticipe l'approche de Jerzy Neyman sur l'inférence *via* des intervalles de confiance. P. M. Boudot a prétendu que Bernoulli était un bon inductiviste et un précurseur des théories de Rudolf Carnap. »

[31] Paul Franceschi, passionné par les paradoxes et notamment les paradoxes probabilistes tels que la Belle au bois dormant, est un des philosophes qui aujourd'hui tentent de réhabiliter Eubulide.

## 2. D'Alembert, le mathé-magicien du temps

Le maître de Condorcet et de Laplace n'est pas resté dans l'histoire pour son « analyse des hasards ». Beaucoup parlent d'un apport négligeable, quelques sévères dénoncent une incroyable errance[32]. Il est néanmoins reconnu aujourd'hui que la lecture attentive des articles, des essais, des justifications de D'Alembert à propos de ses vues sur les probabilités est utile, non seulement pour mieux comprendre ses travaux importants dans d'autres domaines des mathématiques, mais aussi pour mieux comprendre les théoriciens des probabilités qui ont attaqué ces textes. L'encyclopédiste ne voulait qu'y exprimer des doutes, poser des questions, formuler une critique bienveillante : les lois et les calculs que ses contemporains adoptaient ne le satisfaisant pas complètement, il le faisait savoir en argumentant du mieux qu'il pouvait[33]. Cependant, il ne se montra ni suffisamment clair ni suffisamment prudent lorsque, à plusieurs reprises, il proposa une analyse délicate du jeu de pile ou face. Cette analyse, que nous allons tenter de reconstituer, va nous occuper un certain temps parce que nous croyons qu'elle présente des éléments très liés : une distinction instructive entre une « probabilité physique » et une « probabilité métaphysique » et des similitudes troublantes avec les analyses des problèmes d'auto-localisation. C'est une enquête partiellement inédite, risquée mais pleine de promesses qui nous attend.

[32] Diderot et les propres élèves de D'Alembert furent ses premiers critiques. Un siècle plus tard, rien ne s'était arrangé. Cf. Bertrand (1889), préface : « L'esprit de d'Alembert, habituellement juste et fin, déraisonnait complètement sur le Calcul des probabilités » ; Delannoy (1895) : « Il faut reconnaître que d'Alembert s'est complètement trompé, et cela n'a rien d'étonnant, car, tout grand mathématicien qu'il fût, il n'entendait rien aux questions de probabilités. »

[33] D'après Hacking (1971), le collaborateur de Diderot est peut-être même « *the greatest of sceptics about probability mathematics* ».

## 2.1. Le jeu de pile ou face

En 1754 parait le quatrième volume de l'*Encyclopédie*. Une des nombreuses contributions de D'Alembert, l'article délibérément provocateur[34] « Croix ou pile », expose notamment un raisonnement court et hardi qui, aux yeux des mathématiciens amateurs comme des plus aguerris, est pour le moins très suspect de prime abord :

> On demande combien il y a à parier qu'on amènera croix en jouant deux coups consécutifs. La réponse qu'on trouvera dans tous les auteurs, et suivant les principes ordinaires, est celle-ci : Il y a quatre combinaisons : croix (premier coup), croix (second coup) ; pile, croix ; croix, pile ; pile, pile. De ces quatre combinaisons une seule fait perdre, et trois font gagner ; il y a donc 3 contre 1 à parier en faveur du joueur qui jette la pièce. […] Cependant cela est-il bien exact ? Car pour ne prendre ici que le cas de deux coups, ne faut-il pas réduire à une les deux combinaisons qui donnent croix au premier coup ? Car dès qu'une fois croix est venu, le jeu est fini, et le second coup est compté pour rien. Ainsi il n'y a proprement que trois combinaisons de possibles : croix, premier coup ; pile, croix, premier et second coup ; pile, pile, premier et second coup. Donc il n'y a que 2 contre 1 à parier. […] Ceci est digne, ce me semble, de l'attention des calculateurs, et irait à réformer bien des règles unanimement reçues sur les jeux de hasard.

D'Alembert étend son raisonnement à des variantes du jeu plus longues, où face (croix) peut être obtenu en plus de deux lancers de la pièce de monnaie. Néanmoins, nous allons nous contenter de cette version en deux coups pour une raison très simple : il se trouve qu'un pile ou face en deux coups est l'ossature d'un scénario alternatif de la Belle au bois dormant, variation importante (nous allons en parler et en reparlerons) récemment introduite par le jeune philosophe américain Michael Titelbaum, lequel ne fait cependant aucun lien avec D'Alembert, qu'il n'a peut-être jamais lu.

[34] Crépel (2009).

L'article « Croix ou pile » n'énonce évidemment pas un problème d'auto-localisation, son auteur n'a même jamais songé à attribuer une probabilité à une position, encore moins à *sa* position dans l'espace ou le temps. Il réfléchit simplement à l'art de conjecturer les effets d'un dispositif aléatoire, par l'évaluation rationnelle du poids de chaque possibilité, ici chacune des issues d'un ou de plusieurs tirages à pile ou face. Il insiste sur le fait, tout à fait concevable, qu'il n'y a proprement que trois issues possibles au jeu en deux coups, bien que l'on dénombre sans difficulté non pas trois mais quatre combinaisons des orientations de la pièce : face-face, face-pile, pile-face et pile-pile. Il faut en effet confondre les deux premières combinaisons puisque, si face est obtenu du premier coup, un deuxième jet de la pièce est inutile, ne fait pas partie du jeu, lequel s'est terminé par la victoire du joueur qui a parié sur face. Très bien. Ce que nous n'acceptons pas, c'est que l'intuitive cote de 3 contre 1 (soit une probabilité de 1/4 de ne pas obtenir face en deux coups) soit remplacée par une cote de 2 contre 1 (probabilité de 1/3), et cela apparemment en raison d'une étrange conservation de l'équiprobabilité des issues possibles, qu'elles soient trois ou quatre. Nous pensons avec raison que la cote de pile-pile ne peut pas changer lorsque nous n'envisageons plus que trois issues. La probabilité d'obtenir face du premier coup est 1/2 : c'est la somme des probabilités des deux issues face-face et face-pile réduites à une seule, et le double de la probabilité de pile-pile, qui est toujours 1/4. C'est pourquoi nous nous alarmons : non, les trois issues retenues dans le raisonnement concurrent ne sont pas équiprobables, ce raisonnement est biaisé. Puis nous essayons de nous rassurer : D'Alembert veut peut-être juste signifier qu'il existe un raisonnement menant à une cote contre-intuitive de 2 contre 1, que ce raisonnement n'est pas valide mais qu'identifier l'erreur n'est pas aussi facile qu'on le croit. Alors nous relisons l'article, nous lisons d'autres textes de l'auteur, y compris des réponses adressées à ses détracteurs, et nous nous inquiétons à nouveau : il considère bel et bien que le raisonnement du 2 contre 1 entre en

concurrence serrée avec celui du 3 contre 1, et même, à titre personnel, il le préfère, lui trouve plus de sens, plus de cohérence, et finalement, à côté de nombreuses précautions oratoires, n'hésite pas à appeler « paralogisme »[35] le raisonnement du 3 contre 1 ! Il faut donc chercher cette cohérence exotique.

### 2.2. Explication de la position d'alembertienne

Pour se justifier, D'Alembert est amené à considérer deux sortes de probabilités. Les *probabilités métaphysiques* ou *mathématiques* sont notamment attachées à ce que les mathématiciens appellent des combinaisons, c'est-à-dire des associations d'idées très souples, qui notamment autorisent un rangement de certains éléments dans un ordre différent ; par exemple, face-pile et son inversion pile-face sont deux arrangements qui prennent autant de sens l'un que l'autre dans les énoncés où on les place et qui ont même probabilité, ils sont au fond la même combinaison. Les *probabilités physiques* concernent les événements du monde, événements complexes qui sont évidemment idéalisés dans les combinaisons pures ; par exemple, deux tirages consécutifs d'une pièce de monnaie amènent tantôt pile puis face, tantôt face puis face, tantôt pile puis pile, tantôt face puis pile, ce qui donne lieu à quatre combinaisons. Pourtant la théorie mathématique ne reflète pas exactement ce qui se passe dans la nature ; le « cours » de celle-ci, autrement dit le *temps*, est mis en cause :

> C'est qu'il faut distinguer entre ce qui est *métaphysiquement* possible, et ce qui est possible *physiquement*. Dans la première classe sont toutes les choses dont l'existence n'a rien d'absurde ; dans la seconde sont toutes celles dont l'existence non seulement n'a rien d'absurde, mais même rien de trop extraordinaire, et qui ne soit dans le cours journalier des événements. […] Dans le cours ordinaire de la nature, le même événement (quel qu'il soit) arrive assez

[35] D'Alembert (1761), p. 16.

> rarement deux fois de suite, plus rarement trois et quatre fois, et jamais cent fois consécutives […].[36]

Le temps ne s'écoule pas dans le monde des idées, où s'enchainent tout au plus des étapes logiques. Le temps est physique, c'est-à-dire qu'il est l'élément dans lequel les choses naturelles se transforment. Il n'est pas seulement cette idée invariable qui nait du mouvement, du passage des objets physiques, il est presque lui-même un objet physique, pas comme les autres, mais une force, un poids entrainant des transformations qui sont avant tout des alternances du même et du différent. Ce n'est donc pas le temps de beaucoup de savants contemporains de D'Alembert, ce n'est pas la durée sans borne qui permet à tout événement même très improbable de se produire un jour ou, comme l'écrit Diderot, qui « tend à chaque instant à donner une valeur infinie aux quantités finies les plus petites »[37]. La nature est patiente pour la plupart des mathématiciens des Lumières, pas pour D'Alembert. Chez celui-ci, le temps précipite d'une part le remplacement du même par son autre, et empêche d'autre part la venue des événements les plus improbables, tels qu'une pièce qui tomberait sur pile cent fois de suite[38]. Ce temps est en quelque sorte une courbure du temps de Diderot, et sa description fait peut-être écho à des thèses antiques comme l'alternance des opposés dans les fragments du *De la nature* d'Héraclite.

D'Alembert ne refuse absolument pas les théories métaphysiques et les calculs mathématiques de son époque touchant les probabilités. Hors du monde de la causalité et hors du temps, les combinaisons n'ont pas d'histoire. Les probabilités que les mathématiciens leur attribuent sont exactes. Une erreur de ces derniers serait, lorsqu'ils passent de la théorie à la pratique, de ne pas prendre en compte la dépendance des événements

[36] *Ibid.*, p. 10.

[37] Diderot (1761).

[38] La section 3.9 de Paty (1988) décrit ce « temps physique » et ses effets.

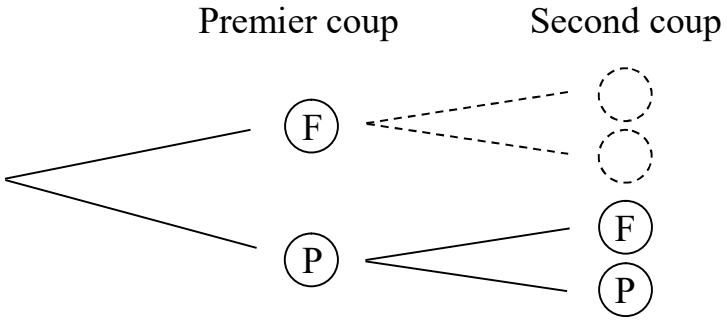
séparés dans le temps, par exemple de croire que le résultat d'un tirage à pile ou face est indépendant des tirages qui l'ont précédé. Une autre erreur, proche de la première, est de ne pas tirer les conséquences du fait que le temps transforme une possibilité en certitude. Ainsi les mathématiciens, qui ont raison de rendre la probabilité (métaphysique) d'une combinaison égale au produit des probabilités (métaphysiques) des éléments constituants, se trompent dès qu'ils appliquent ce calcul à des événements tels que la venue de l'un dépend de la venue d'un autre. Par exemple, penser que la probabilité (physique) de pile-pile est 1/2.1/2, c'est composer des probabilités (physiques) de natures différentes, une probabilité qui suppose qu'un second tirage doit avoir lieu et la probabilité qu'il ait lieu : c'est, de façon contradictoire, considérer le tirage à la fois comme nécessaire et comme seulement possible[39].

Reprenons. Il y a trois issues au jeu en deux coups : face, pile-face et pile-pile. Il n'est pas permis de calculer les probabilités des deux dernières comme on calcule les probabilités de combinaisons idéales. Il ne reste plus qu'à respecter un principe d'indifférence[40]. Mais on peut rétorquer qu'il n'est pas possible, pour un seul tirage-face, d'être aussi probable qu'un tirage-pile suivi d'un tirage-face. Il y a un complet déséquilibre entre les trois hypothèses « le premier lancer donne face », « le premier lancer donne pile, le second donne face » et « les deux lancers donnent pile ». D'Alembert aurait du mal à répondre à cette objection, mais il y répondrait sûrement, et il est même envisageable de trouver la piste d'une solution

[39] D'Alembert (1761), p. 18.

[40] Dans des écrits ultérieurs, D'Alembert cherche à calculer les probabilités physiques de face, pile-face et pile-pile, notamment à rendre pile-face un peu plus probable que pile-pile qui est une répétition du même. Cependant, notre explication générale n'est pas précisément une justification de l'équiprobabilité, mais une justification de la préférence d'une équiprobabilité de trois combinaisons à une équiprobabilité de quatre combinaisons. Au fond, ces calculs ultérieurs ne sont pas incompatibles avec notre explication, et nous ne les détaillons pas parce que nous n'en avons pas besoin.

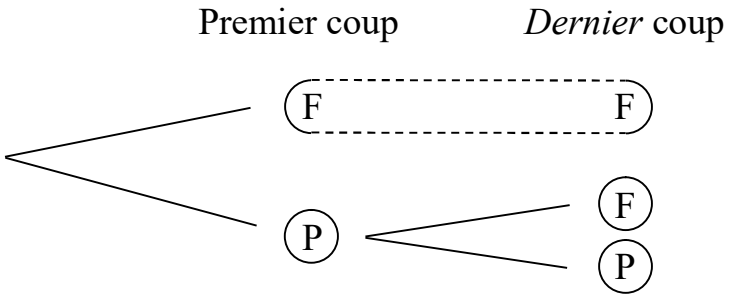
dans son texte. L'extrait à se passer et se repasser dans la tête, en soulignant les mots importants, est sans doute celui-ci : « ne faut-il pas *réduire à une les deux* combinaisons qui donnent croix au *premier coup* ? Car dès qu'une fois croix est venu, *le jeu est fini*, et le *second coup est compté pour rien* ». Nous nous rappelons aussi que D'Alembert insiste sur le fait qu'un jeu est toujours *fini* dans des délais raisonnables, ou plutôt dans des délais naturels : il est persuadé que des lancers successifs de la pièce ne peuvent pas s'éterniser, face doit forcément arriver et même se hâte, la nature ayant horreur des longues répétitions du même. Le second coup n'est rien[41], ce n'est même pas une trace blanche sur le schéma qui pourrait représenter le déroulement du jeu et les tirages possibles, parce que dans cet espace on pourrait encore placer un second coup virtuel :




En fait, il n'est pas légitime de parler de « second coup » puisque celui-ci n'est pas assuré de se produire. Dans le temps du jeu, il y a en revanche un premier coup et un dernier coup, il y a, non pas le temps absolu du premier coup et le temps absolu du second, mais deux temps relatifs aux limites du jeu : le temps de début et le temps final. Il y a possibilité pour face d'être à la fois premier et dernier coup, ce qui signifie que le temps de début et le temps de fin se confondent. La relativité du temps selon D'Alembert, c'est la possibilité d'*un repli à l'extrême de deux positions temporelles*, et plus seulement d'un léger pli en accordéon ou

[41] Hacking (1971) met en scène le jeu de croix ou pile en précisant, pour renforcer l'idée d'un anéantissement du second coup, que la pièce est jetée dans un four dès que le jeu est fini. On pense encore au puits sans fond ou à d'autres métaphores.

d’une courbure. Le jeu a trois issues possibles de même nature, énoncées par trois propositions équilibrées : « face arrive au premier coup et face au dernier coup » ; « pile arrive au premier coup et face au dernier coup » ; « pile arrive au premier coup et pile au dernier coup ». Le schéma que D’Alembert a en tête serait donc celui où le premier coup et l’autre coup (dernier et non second) peuvent être un seul et même coup, celui où en quelque sorte le temps ne s’écoule plus quand face est venu :




Voilà qui justifie enfin l’application du principe d’indifférence et l’équiprobabilité des issues… quoiqu’il reste un souci. Le premier tirage face occupe toute la place, tout le temps du jeu, car il est aussi premier et dernier. En tant que premier, il a même probabilité que pile, c’est-à-dire 1/2, mais en tant que premier et dernier, il a même probabilité que pile-face et pile-pile, soit 1/3. Cela n’est-il pas contradictoire ? Les règles du jeu sont telles que si face arrive au premier coup, alors face arrive au premier et au dernier coup, et inversement : on est tenté de dire qu’en raison d’une équivalence entre F = « Face arrive au premier coup » et FD = « Face arrive au premier coup et face arrive au dernier coup », la probabilité de face au premier coup devrait être la probabilité de face aux premier et dernier coups. Or on a deux probabilités différentes, 1/2 et 1/3, ce qui étonne pour le moins. Si l’on considère que FD est la conjonction de F et d’une proposition qui décrit un événement dont la date est relative à la fin du jeu (« Le *dernier* coup donne face »), alors il est possible que l’équivalence de F et FD ne soit pas une équivalence logique habituelle…

Gardons cela dans un coin de notre mémoire, nous verrons bientôt si une explication plus complète est envisageable.

### 2.3. La Belle au bois dormant s'en mêle

Portons à présent un regard attentif sur le problème de la Belle au bois dormant, la fameuse énigme d'auto-localisation que nous avons énoncée dans l'introduction de cette thèse. Là aussi, le temps nous joue des tours. Dans leurs arguments les plus spontanés, les pro-1/2 ou demistes, c'est-à-dire les chercheurs qui pensent qu'une Belle incapable de dater son réveil doit croire au degré 1/2 que la pièce est tombée sur face, se focalisent sur une pièce équilibrée : dès le dimanche, la veille de l'expérience, ils alignent le degré de croyance en face sur la propension de la pièce à tomber sur face, puis ils conservent ce degré quand la Belle se réveille dans l'expérience, en arguant qu'elle ne reçoit aucune information susceptible de le modifier. Les pro-1/3 (tiéristes) se concentrent quant à eux sur le reste du dispositif aléatoire, sur les réveils, qui vont par deux en cas de pile et seulement en cas de pile : ils pensent que la Belle engagée dans l'expérience doit considérer des proportions ou des fréquences de réveils, constater qu'il y a deux fois plus de réveils-pile que de réveils-face, et ce n'est qu'ensuite que la Belle doit croire au degré 1/3 que la pièce est tombée sur face par alignement sur la proportion des réveils-face. Par conséquent, on peut dire que les pro-1/3, par rapport aux pro-1/2, retardent l'alignement entre probabilités ontique et épistémique. Pour la Belle qui subit une perte de repère temporel à cause de la drogue à effet amnésique, il n'y a pas qu'une hésitation entre lundi et mardi, il y a en quelque sorte superposition d'un « temps demiste » où l'agent hésite, comme le dimanche, entre face et pile sans se préoccuper de la date du réveil, et un « temps tiériste » où l'agent hésite entre trois possibilités : face-lundi, pile-lundi et pile-mardi. Ne manquons pas de remarquer que face-mardi est interdit par le protocole : le fait que l'agent reste inconscient le mardi en

cas de face change les estimations tiéristes ; ce saut de lundi à mercredi, ce mardi englouti qui n'est que néant pour la conscience de l'agent est aussi important que le « rien » qui remplace le second coup du « croix ou pile » d'alembertien.

Reprenons. Les tiéristes, majoritaires dans la littérature, déforment l'espace de probabilité de la Belle : elle voit pile en plus gros, elle est passée d'une situation où sa position temporelle est sans importance à une situation où elle est pertinente pour l'issue pile/face. Cela rappelle beaucoup l'analyse de D'Alembert qui dit en gros : lorsque vous conjecturez l'avenir, si vous vous situez dans le temps d'un lancer de pièce singulier, alors la probabilité de face (premier coup) est 1/2 ; en revanche, si vous vous situez dans le temps total du jeu en deux coups, la probabilité de face (premier et dernier coup) passe à 1/3. Et il y a mieux : les analyses du paradoxe de la Belle au bois dormant sont elles aussi confrontées à une équivalence logique particulière, celle qui ferait correspondre les propositions F = « La pièce tombe sur face » et FL = « La pièce tombe sur face et aujourd'hui est lundi ». FL est « plus » que F, c'est la conjonction de F et de « Aujourd'hui est lundi » qui est en outre une proposition indexicale[42] dont la valeur de vérité peut changer avec le temps puisqu'elle dépend évidemment du jour qu'on est. Par conséquent, il ne peut pas y avoir une équivalence logique stricte entre F et FL. Mais pour le sujet Belle engagée dans l'expérience qui interdit les réveils-face-mardi, F est vraie si et seulement si FL est vraie : il y a donc équivalence *dans le contexte de l'expérience*, c'est-à-dire pour le sujet de l'expérience en cours. Élevons en principe l'équiprobabilité des propositions logiquement équivalentes. Quand la plupart des chercheurs qui ont travaillé sur la Belle au bois dormant disent qu'au cours de l'expérience la Belle doit croire que F et doit

[42] Un indexical de temps est un mot ou une expression qui indique un moment relatif au (temps du) sujet qui le prononce. « Aujourd'hui est lundi » contient l'indexical « aujourd'hui ».

croire que FL au même degré, ils ne suivent pas exactement ce principe mais un principe étendu aux propositions *équivalentes dans un certain contexte*. Faut-il vraiment accepter le principe étendu ? La réponse à cette question n'est pas si simple. Un auteur[43] a fait remarquer qu'une Belle qui ignore si on est lundi ou mardi, qui ne peut pas se repérer dans le temps au jour près, n'est peut-être pas tenue d'associer à F, proposition sans indexical de temps, la même probabilité épistémique qu'à FL, proposition clairement indexicale. Une probabilité de 1/2 pour F et une probabilité de 1/3 seulement pour FL ne seraient pas irrationnelles. Nous aurons l'occasion d'approfondir cela dans le quatrième chapitre.

Sur le point précis de l'équivalence, l'analogie avec le pile ou face de D'Alembert est difficile. Entre « Face arrive au premier coup » et « Face arrive au premier coup et face arrive au dernier coup », il n'est pas clair que l'équivalence soit seulement contextuelle étant donné que, dans la seconde proposition, les deux termes de la conjonction apparaissent eux-mêmes équivalents. Dans « La pièce tombe sur face et aujourd'hui est lundi », le deuxième terme de la conjonction donne une position temporelle et pas du tout le résultat d'un tirage à pile ou face. Comment réparer l'analogie ? Peut-être en examinant, comme annoncé quelques pages plus haut, un scénario alternatif de l'expérience de la Belle récemment proposé dans la littérature[44]. Notre énoncé conserve à peu près la structure de la variante, il ajoute surtout un certain personnage célèbre :

La Belle et D'Alembert :

La Belle accepte de participer à une nouvelle expérience organisée par le grand D'Alembert. Elle en connait toutes les règles. Le lundi, le mathématicien la réveille, a un entretien avec elle, la rendort avec la drogue

[43] Peter Lewis (2010).

[44] Cf. Titelbaum (2012), p. 148.

à effet amnésique, puis lance une pièce de monnaie équilibrée. Si face, il la laisse dormir mardi. Si pile, il la réveille mardi, a un entretien avec elle, la rendort avec la même drogue, puis relance la pièce ; le résultat de ce second tirage est sans conséquence. Dans tous les cas, la Belle est réveillée mercredi et son aventure est terminée.

Cette expérience alternative ressemble à l'originale, on ne compte que deux modifications structurelles : un déplacement, du dimanche au lundi, du tirage qui décide si la Belle sera réveillée mardi ; un second tirage le mardi si et seulement si pile est venu au premier tirage (si et seulement si la Belle est réveillée mardi). *A priori*, ces différences ne vont rien changer au débat entre les demistes et les tiéristes si celui-ci est comme d'habitude centré sur le tirage qui décide ce qui se passe mardi. En quoi consiste cependant cette expérience organisée par D'Alembert ? C'est en somme un pile ou face en deux coups pimenté par les réveils de la Belle, laquelle, en quelque sorte, suit le destin de la pièce de monnaie « anéantie » le mardi en cas de face au premier coup. Les propositions d'alembertiennes décrivant les issues possibles du *premier* coup et éventuellement du *dernier* coup ont maintenant un sens pour la Belle, qui n'est peut-être pas disposée, lors de l'entretien ou des entretiens prévus avec l'organisateur, à accorder à F = « Face arrive au premier coup » la même probabilité qu'à FD = « Face arrive au premier coup et face au dernier coup ». En effet, F est clairement une possibilité exprimée par une Belle demiste non préoccupée par sa localisation temporelle. Pour que le tiérisme accorde à F la même probabilité qu'à la proposition indexicale qu'il préfère, à savoir FL = « Face arrive au premier coup et on est lundi », il doit accepter le principe d'équiprobabilité des propositions contextuellement équivalentes. Si le principe est refusé, la probabilité 1/2 s'impose pour F comme pour son autre, « Pile arrive au premier coup ». En revanche, FD est une des trois issues possibles du pile ou face en deux coups, avec « Pile arrive au premier coup et face au dernier coup » et « Pile arrive au premier coup et

pile au dernier coup » : pour penser ce trio, il est apparemment nécessaire de prendre en compte l'entièreté du dispositif aléatoire et même de fonder les probabilités épistémiques sur des fréquences de réveils dans une longue série d'expériences répétées ; c'est un procédé similaire au procédé tiériste de comptage des réveils face-lundi, pile-lundi et pile-mardi. FD et FL présentent, certes, les différences logiques que nous avons mentionnées plus haut, et pourtant nous intuitionnons leur probabilité 1/3 de la même manière. Si la Belle est à l'écoute de ces quelques réflexions, il lui parait rationnel d'attribuer aux propositions d'alembertiennes F et FD des degrés de croyance différents, à savoir 1/2 et 1/3 respectivement. Nous savons qu'il s'agit des probabilités « physiques » que D'Alembert lui-même assignerait dimanche, la veille de l'expérience, aux événements face en tant que premier coup et face en tant que premier coup et dernier coup, en suivant les principes de son art de conjecturer l'avenir et alors que l'auto-localisation n'est même pas une vague idée perdue dans les réflexions de ses contemporains occupés à prier le dieu Bernoulli.

### 2.4. Retour sur les deux probabilités d'alembertiennes

Maintenant que nous avons expliqué, notamment grâce à une comparaison anachronique hardie avec un problème d'auto-localisation, les surprenantes estimations probabilistes de D'Alembert, revenons sur les probabilités « physiques » et « métaphysiques ». Selon D'Alembert, la probabilité métaphysique de face est 1/2, que face représente dans le monde intelligible l'événement sensible que constitue une pièce équilibrée qui tombe sur face lors d'un lancer isolé (et non sur pile, unique autre possibilité), ou bien cet événement qui est une des trois issues possibles du jeu en deux coups (face est alors, dans la pensée des mathématiciens, gros des combinaisons face-pile et face-face, ce qui explique qu'il soit deux fois plus probable que pile-pile ou que pile-face, les représentations des deux autres issues). La probabilité physique du premier événement précité est

alignée sur la probabilité métaphysique de face, car c'est un événement élémentaire, qu'on ne peut donc pas décomposer en événements qui se succèdent dans le temps. Par contre, le second événement (face en tant que premier coup et dernier coup) ne doit pas être considéré comme un événement élémentaire mais comme un double événement replié par la fin précipitée du jeu : il n'est pas étonnant, aux yeux de D'Alembert, que sa probabilité (physique) s'écarte de la probabilité (métaphysique) de son idéalisation qui concentre deux combinaisons, face-pile et face-face. Ainsi l'encyclopédiste, qui n'explicite pas complètement la nature de ces deux genres de probabilités, s'investit néanmoins dans le problème de leur alignement. Vraisemblablement, son intérêt parfois naïf pour les énigmes probabilistes de son époque, telles que le paradoxe de Saint-Pétersbourg et, bien sûr, le jeu de croix ou pile (qu'il est à peu près le seul à considérer comme polémique), l'a conduit à distinguer les deux genres.

Mais une objection voudrait nous interdire de penser que la distinction d'alembertienne est perspicace : D'Alembert n'oppose pas un concept épistémique et un concept ontique, il décrit en réalité un seul concept ontique proche de la chance, mesure de la possibilité des événements. La probabilité métaphysique, qui complète la probabilité physique, n'est pas du tout comparable à un degré de croyance ou de confirmation, mais est simplement la plus objective des probabilités, celle des combinaisons mathématiques, des événements idéalisés dont les lois sont plus simples et plus connues que les lois de la nature. Cette objection, qui dit plusieurs fois le vrai, repose néanmoins sur des incompréhensions et des présupposés erronés. D'Alembert n'emploie pas du tout le mot « physique » dans le sens que l'on attendrait aujourd'hui, personne ne prétend que « physique » qualifie la probabilité ontique ou que « métaphysique » qualifie la probabilité épistémique, et même personne n'affirme que D'Alembert sépare deux *concepts*.

Mal interprété, le mot « physique » évoque des idées inappropriées comme « objectif » ou « empirique » et altère aussi le sens de « métaphysique » ; c'est un leurre redoutable. En réalité, la probabilité métaphysique est la probabilité objective de combinaisons, issue d'un calcul infaillible qui ne dépasse pas le cadre de la théorie mathématique pure ; en tant que propriété des êtres du monde intelligible, en quelque sorte elle seule peut être qualifiée d'« ontique ». La probabilité physique est la probabilité *attendue* ou *estimée* des événements du monde sensible, qui incarnent de plusieurs manières les combinaisons ; elle résulte principalement d'un exercice du jugement plus poussé mais paradoxalement plus subjectif, une réflexion qui essaie de prendre en compte les lois de la nature, *lesquelles ne sont pas toutes connues*. La probabilité physique ne peut donc pas être considérée comme « simplement empirique », et d'ailleurs les vérifications expérimentales ne sont pas le souci de D'Alembert, qui préfère les expériences de pensée. La probabilité 1/3 de face en tant que premier et dernier coup est absurde dans une interprétation simplement empirique, seul le cadre épistémique classique la rend acceptable : si nous sommes également indécis par rapport aux trois cas de la partition {face(-face), pile-face, pile-pile}, nous postulons l'équiprobabilité. Nous avons tout à l'heure fait correspondre des degrés de croyance recommandés par les analystes du problème de la Belle au bois dormant avec des estimations physiques de D'Alembert ; cela n'est pas choquant si nous percevons, derrière le masque trompeur du mot « physique », un visage épistémique. La probabilité physique est une correction de la probabilité métaphysique grâce à des réflexions d'ordre épistémologique, orientées, il est vrai, par une certaine expérience du monde de la causalité et du temps ; une probabilité physique non nulle est assignée aux événements, aux choses « dont l'existence n'a rien d'absurde », c'est-à-dire « métaphysiquement possibles », mais en plus dont l'existence est suffisamment croyable et suffisamment constatée dans le « cours ordinaire de la nature ».

### 2.5. Critique de la position d'alembertienne

Nous avons révérencieusement cherché de la cohérence chez D'Alembert, mais préparons-nous à durcir le ton car cette recherche a peut-être une limite. Il est certain qu'aucun joueur n'accepte l'argument du pari à 2 contre 1… sauf pour escroquer un joueur débutant, éventuellement. C'est bien le point faible de D'Alembert : on dirait qu'il ne vit pas dans notre monde, alors qu'il parle de probabilités physiques et de paris, et ses expériences de pensée ne peuvent pas réellement éprouver sa théorie. Que nous dit-il au fond ? Que la probabilité d'obtenir face en un coup avec une pièce de monnaie dépend du nombre de lancers supplémentaires que nous envisageons de faire en cas d'échec, donc dépend d'un certain protocole, pour ne pas dire cérémonial ! Que des événements successifs qu'après examen tout le monde juge indépendants subissent pourtant les transformations dont seuls les penseurs antiques ont le secret ! Que dans la nature il y a des lois cachées, peut-être même à jamais, donc des lois occultes, qui expliqueraient que les probabilités ne sont pas celles que les mathématiciens attendent ! N'ayons pas peur de l'affirmer : il s'agit là d'un résidu de pensée magique. D'Alembert voit les événements du monde se plier, pour ainsi dire, aux prédictions de puissants « maîtres en probabilités ». Comprenons l'intérêt immense des écrits d'alembertiens, admettons avec les historiens que l'article « Croix ou pile » expose pour la première fois un paradoxe du partitionnement des (équi)possibilités qui défie la théorie des probabilités ; en même temps, ne nous voilons pas la face devant cette pensée magique qui, certes, est loin d'avoir corrompu le philosophe, mais qui est perceptible. Sans elle, D'Alembert (le vrai, pas le génie lisse que nous pourrions fantasmer) n'aurait jamais rédigé « Croix ou pile ».

Pourquoi relever cette tentation obscurantiste ? Pour trois raisons. La première est qu'on le fait pour des penseurs très respectés, et sans mauvaise

intention ; c'est peut-être même le signe qu'on les respecte vraiment[45]. La deuxième est que cette tentation est aussi la nôtre, notamment quand nous nous attardons sur les paradoxes probabilistes (nous en apporterons des preuves plus tard). La troisième est que, si on ne le fait pas, des esprits malhonnêtes ou crédules s'en chargeront, mais à leur manière. La philosophie des Lumières a combattu l'obscurantisme sans l'éradiquer, celui-ci s'est adapté, a trouvé de nouvelles stratégies de conquête. Déjà, vers la fin du 18e siècle, tel un faux prophète qui chercherait à pervertir l'esprit des Lumières, le charlatan Cagliostro, qui prédisait les numéros de la loterie et séduisait les naïfs grâce à son apparent contrôle des aléas, évoquait et faisait témoigner les fantômes de D'Alembert et d'autres philosophes appréciés récemment décédés[46]. Quelques superstitieux ont attribué à D'Alembert une martingale pour gagner à la roulette ou à d'autres jeux de hasard ; elle consiste à miser, par exemple, sur rouge ou sur noir, et en cas d'échec de parier à nouveau *sur la même couleur*, mais en doublant la mise. Reconnaissons que l'attribution est pertinente si l'on considère que D'Alembert partage avec beaucoup de joueurs le sentiment ou le préjugé selon lequel le différent a plus de chances de se produire que le même.

Cela étant dit, ce n'est peut-être pas notre monde aux lois immuables qui fait échouer l'analyse d'alembertienne, mais la façon dont on le perçoit. Que fait un magicien ? À l'aide d'une drogue, il se met dans un état de

[45] Il est notoire que Leibniz, dont les méditations précieuses sur les mondes possibles en font un précurseur de Kripke et de Lewis, retrouvait par ailleurs son « arithmétique binaire » dans le *Yi Jing*, le vieux manuel chinois de divination que son ami le jésuite Bouvet considérait comme une des nombreuses inventions d'Hermès Trismégiste.

[46] Tatin-Gourier (1994) explique les buts de Cagliostro dans une préface éclairante et propose les pamphlets et les mémoires du procès de « l'affaire du Collier » qui, entre autres, dénoncent les méthodes du charlatan ; les exagérations de ces « témoins » historiques ont contribué à rendre celui-ci célèbre dans le petit monde assez conspirationniste des « occultistes » contemporains.

conscience particulier, il triche avec les lois physiques qu'il meut de son seul point de vue. Le protocole de la variante de la Belle et D'Alembert apparait comme un rituel nécessaire à la transformation d'une probabilité 1/2 en 1/3, il donne raison au collaborateur de Diderot.

Quand on se penche, même avec un œil critique et sévère, sur les questions posées par D'Alembert aux spécialistes des probabilités, ce sont des merveilles que l'on découvre. Il devance son élève Condorcet sur le discernement d'une double identité de ce curieux concept éclairé un siècle plus tôt par Pascal. Pourtant, il semble davantage tirer sa distinction d'une méditation personnelle sur le temps que du contexte philosophique : il ne pouvait qu'être incompris. Son époque n'était pas prête à accueillir quelque chose comme « le dernier coup », une expression qui travestit en réalité une indexicalité de temps, puisque la date du dernier coup dépend du premier, comme un réveil-mardi dépend du résultat d'un tirage à pile ou face. Mais nous allons entrer dans une autre époque, celle qui voit naitre l'auto-localisation dans des travaux philosophiques parfois proches des réflexions sur la double interprétation de la probabilité.

# Chapitre 2

# L'auto-localisation et David Lewis

David Lewis est sans doute le penseur de l'auto-localisation le plus cité dans les travaux sur le sujet. Il a aussi réfléchi à la nature de la probabilité et à un principe d'alignement ontique-épistémique qu'il appelle le principe principal, des réflexions qui jouent un rôle dans les résolutions des problèmes d'auto-localisation. C'est donc naturellement autour de lui que ce chapitre est construit. Dans un premier temps, nous présentons quelques pensées et théories qui ont participé à l'émergence de la théorie de l'auto-localisation et qui ont influencé Lewis. Dans un deuxième temps, nous regardons de près ses travaux les plus pertinents et analysons déjà quelques problèmes instructifs. Enfin, nous nous tournons vers trois autres théoriciens, John Perry, Nick Bostrom et Robert Stalnaker, pour comprendre leurs spécificités, leurs convergences et leurs divergences.

# 1. Langage et (cosmo)logique

## 1.1. L'indexicalité dans *La Pensée* de Frege

C'est en 1918 que Frege publie *La Pensée* (*Der Gedanke*), première des trois *Recherches logiques*. En 1956, une traduction en anglais parait dans *Mind*. Cette nouvelle publication connait un plus grand succès et le problème, pourtant rapidement évoqué, des démonstratifs et des indexicaux dans le langage ordinaire ne laisse pas indifférent. Rappelons d'abord que, pour Frege, les *pensées* sont les significations « immatérielles » des phrases d'un langage, les contenus objectifs véhiculés par ces phrases, par opposition à leurs occurrences « matérielles », c'est-à-dire écrites ou prononcées ; ce sont les entités structurées d'un monde idéal, qui ont la propriété d'être vraies ou fausses absolument, éternellement, indépendamment des subjectivités qui les jugent. Une phrase peut dire plus que la pensée, ou moins que la pensée. Dans la plupart des cas, elle se charge d'éléments au mieux complémentaires, au pire étrangers : un signe qui la fait reconnaitre comme assertion ou comme question, un mot qui trahit une humeur ou une émotion, ou encore un effet poétique, un rythme, un ton…

> Ainsi n'est-il pas rare que les contenus d'une phrase dépassent la pensée qui y est exprimée. Mais l'inverse se produit tout aussi bien ; il arrive que le simple énoncé verbal, ce que fixent l'écriture et le phonographe, ne suffise pas à l'expression de la pensée. Le *tempus praesens* est employé de deux manières différentes : en premier lieu pour donner une indication de temps, en second lieu pour supprimer toute limitation dans le temps, quand l'intemporalité ou l'éternité est partie intégrante de la pensée. […] Si le *praesens* comporte une indication de temps, il faut savoir quand a été prononcée la proposition pour comprendre correctement la pensée. Le temps où les paroles sont prononcées est alors une partie de l'expression de la pensée. Si on veut dire aujourd'hui la même chose qui fut exprimée hier avec le mot « aujourd'hui », on remplacera ce mot par « hier ». Bien que la pensée soit la même, l'expression verbale doit être

différente, pour compenser la modification de sens que la différence des moments où l'on parle ne manquerait pas de produire. Il en va de même avec des mots comme « ici », « là ». Dans tous les cas semblables, le simple énoncé verbal, tel qu'il peut être fixé par l'écriture, n'est pas l'expression complète de la pensée. Pour la comprendre correctement, il faut connaitre en outre les circonstances précises qui accompagnent les paroles et qui servent à l'expression de la pensée. On peut ajouter les signes du doigt, les gestes, les regards. Un même énoncé où figure le mot « je » exprime des pensées différentes dans la bouche de personnes différentes, et il peut se faire que de ces pensées les unes soient vraies et les autres fausses.[47]

Cette réflexion sur le manque possible d'intégrité ou d'intégralité des phrases ordinaires entendues sans le contexte de leur énonciation met en avant des indexicaux et amène de surcroit le triplet cher aux penseurs de l'auto-localisation : temps, lieu, identité de l'agent. Vigilant quant à la cohérence de son système, Frege doit refuser de rendre les propositions indexicales conformes à des pensées, lesquelles sont soit vraies, soit fausses, propriété qui ne change pas relativement aux conditions de leur « matérialisation » (quand ? où ? qui ?). Tout rapport d'un fait, d'un événement, prétend à l'« éternité » des lois de la nature et des mathématiques, mais il doit incorporer toutes les informations nécessaires et donc bannir les indexicaux, qui invitent à un complément d'information dès que les circonstances de leur expression sont oubliées.

Frege remarque dans la suite du texte, sans être très explicite, que nous avons souvent l'impression de formuler et de communiquer toute notre pensée même en usant de démonstratifs, tant que nous partageons avec des interlocuteurs connus une portion restreinte de l'espace et du temps. C'est d'ailleurs pour cela que nous employons en toute confiance le pronom « je » et que nous ne nous désignons que très rarement à la troisième personne en donnant notre nom ou nos qualités. Malgré tout,

[47] Frege (1971), p. 178.

cette communicabilité doit être considérée comme accidentelle, et les plus hautes pensées ne peuvent être transmises sans erreur aux hommes attentifs que par le moyen de matérialisations diverses et espacées qui évitent les indexicaux imprécis.

Nous donnerons en son temps l'analyse du théoricien de l'auto-localisation John Perry qui, conscient de la gêne éprouvée par Frege, l'explique en montrant qu'en certaines occasions les individus ne peuvent exprimer une pensée qu'à l'aide d'un indexical. Notons que Frege est l'inventeur du personnage de Rudolph (ou Rudolf) Lingens, réutilisé plus tard dans des fictions illustrant l'auto-localisation (Perry, Lewis, Stalnaker) ou seulement l'indexicalité du langage naturel (David Kaplan).

### 1.2. Mondes possibles et chats noirs

Leibniz n'est peut-être pas l'inventeur de l'expression « mondes possibles » au pluriel, mais c'est une autorité qui a séduit les nouveaux métaphysiciens contemporains. Les historiens font du moderne allemand un précurseur, en admirant son traitement de l'idée d'une pluralité d'autres mondes contenus dans un entendement supérieur, qui n'ont aucune existence autre que logique[48]. Lewis témoigne qu'il connait très mal Leibniz et regrette cette ignorance avec humilité[49]. Certes, aujourd'hui les chercheurs ne renvoient plus les difficultés posées par l'étude des mondes à la théologie, comme le faisait le bibliothécaire de la maison de Hanovre.

[48] Les *Essais de Théodicée* sont publiés en 1710. Le Dieu leibnizien est tout-puissant dans le sens où il est capable de concevoir et de faire venir à l'existence tous les mondes possibles qui ne sont pas logiquement contradictoires. Il porte néanmoins à l'existence un seul de ces mondes, non pas parce qu'il a une faiblesse, mais parce qu'il est bon et capable de choisir le meilleur des mondes. Être nécessaire, il est la dernière raison des événements contingents de notre monde, qui est, en vertu du principe de raison suffisante « *nihil est sine ratione* », le seul monde créé.

[49] David Lewis (2007), préface, p. 12-13.

L'expression « mondes possibles » est revenue dans les travaux sur la logique modale, notamment chez l'influent Saul Kripke[50], que Lewis connaissait évidemment beaucoup mieux. La logique modale ajoute des modificateurs qui jouent le rôle des adverbes et des modes grammaticaux en français (des auxiliaires modaux *can*, *must*, *may*, *would*… en anglais). Les longs développements n'étant pas utiles à notre objectif, contentons-nous d'exemples très simples.

Si la proposition « Tous les chats sont noirs » parait fausse, elle peut être modérée ou modulée de plusieurs manières pour un résultat plus crédible, par exemple : « Tous les chats du Japon pourraient être noirs à partir de l'an 2500 ». Ici, on ne fait qu'exprimer la possibilité que la première proposition soit vraie dans une époque future et un lieu restreint : ce déplacement, simplement spatio-temporel malgré la pensée d'une *possibilité*, rend de prime abord contestable que quelque chose comme des mondes alternatifs soit intervenu. Mais que se passe-t-il quand nous supposons, dans un propos qui, vrai ou faux, nous semble en tout cas sensé et légitime, que « si les chats n'avaient pas de moustaches, alors ils se cogneraient plus souvent » ? Nous ne couperons pas les moustaches des pauvres félins ! Cette fois-ci, le déplacement n'est pas spatio-temporel : nous désignons clairement un possible monde proche du nôtre, similaire, mais dans cet autre univers qui ne reflète pas exactement notre commune chaine des causes et des effets, les chats n'ont pas de moustaches. Ces chats-là sont des *contreparties*, des alternatives aux chats de notre monde, et ce qui leur arrive et que nous associons à des jugements vrais ou faux sont des *contrefactuels*, des alternatives aux faits de notre monde. On appelle d'ailleurs *conditionnel contrefactuel* le mode qui ouvre la voie vers un autre monde ; c'est aussi le nom donné à un énoncé conditionnel (de

[50] Voir surtout ses articles sur la logique modale publiés entre 1959 et 1963.

type « si… alors… ») lorsqu'il emploie ce mode, ce qui est le cas de notre exemple.

Revenons maintenant sur la première difficulté : nous accordons raisonnablement à « Tous les chats du Japon seront noirs à partir de l'an 2500 » une probabilité très faible mais non nulle, ce qui justifie l'emploi du verbe pouvoir au conditionnel dans « Tous les chats du Japon pourraient être noirs à partir de l'an 2500 ». Bien qu'enclins à croire que les chats nippons ne vont pas perdre la diversité de leurs couleurs, nous admettons que nos connaissances limitées ne nous permettent pas d'en être absolument sûrs, que l'invraisemblable peut advenir suivant des causes ignorées. C'est que nous nous projetons dans des mondes possibles pour évaluer leur adéquation avec le nôtre. Des mondes nous sont épistémiquement accessibles : ce sont tous ceux qui pourraient être notre monde en raison de tout ce que nous savons. Parmi eux, des mondes nous semblent moins accessibles, moins possibles, nous les saisissons dans des attitudes modérées que le langage ordinaire rend notamment par le conditionnel. Nous admettons qu'au moins un monde accessible présente un Japon futuriste où tous les chats sont noirs, il pourrait être notre monde.

Ce qui met d'accord les nombreux théoriciens des mondes possibles, c'est que ces mondes ne sont pas des fantaisies, des produits inutiles de l'imagination, mais ils fournissent une structure de référence grâce à laquelle nous caractérisons notre monde. Remarquons que la science use souvent de falsifications délibérées où des lois physiques sont abolies, où des difficultés sont éliminées. Pour étudier par exemple le déplacement d'un mobile sur un plan et calculer sa position à un instant donné, nous choisissons de « négliger » les frottements, mais à bien y penser, ce que nous faisons, c'est que nous nous représentons un monde fictionnel où les frottements n'existent pas et nous examinons comment s'y déplace la

contrepartie de notre mobile[51]. Quand nous disons que cette fiction est proche de la vérité quant à ce qui se produit réellement sur le verglas, nous parlons en bon physicien. Quand nous disons en quoi des phénomènes complexes de notre monde ressemblent à des idéalisations plus simples, nous approchons, voire nous exprimons une vérité sur notre monde, une vérité que nous ne pourrions admettre ou faire admettre que difficilement si notre esprit était incapable d'idéaliser de la sorte. Nous avons peut-être peu d'occasions d'en témoigner, et pourtant nous comprenons et effectuons souvent des comparaisons entre les autres mondes et notre monde afin de connaitre ce dernier.

### 1.3. Quine et les attitudes égocentriques

Nous pénétrons l'intimité des mondes possibles quand nous croyons telle ou telle proposition mais aussi quand nous voulons, craignons, regrettons, quand nous avons diverses *attitudes* par rapport aux propositions. Cette réflexion est incorporée par W. V. O. Quine dans une analyse sur l'équivocité des phrases ordinaires, qui peuvent avoir, notamment à cause des indexicaux, des significations différentes (propositions logiques et non pensées comme chez Frege) et ainsi être vraies à certains moments, fausses à d'autres. En 1965, le maître de David Lewis se rend dans plusieurs universités américaines pour donner la lecture de son « essai » final, *Propositional Objects*, texte important où l'émergence des concepts d'attitude *de se* et de croyance auto-localisante est particulièrement sensible. Pourtant, la principale illustration utilisée et réutilisée dans ces conférences, et que nous reprenons ici, est assez singulière puisqu'un chat en est le héros, considéré comme un individu en attitude, entretenant un rapport avec des propositions et des mondes. Malgré les précautions prises par l'auteur (pour qui un animal, au langage

[51] Nous empruntons cet exemple à David Lewis (2007), p. 52.

trop limité ou inexistant, ne peut certainement pas avoir des croyances mais, au moins, a une volonté, fonction strictement physiologique), le résultat est peut-être plus déstabilisant que la description plaisante de chats futuristes ou même que la mention (que faisait Lewis) de vaches qui volent dans d'autres mondes.

Un chat cherche à grimper sur un toit pour échapper à un chien. Ce que le chat veut, c'est l'état de choses qui est la classe de tous les mondes possibles dans lesquels il atteint le toit, alors que la classe de tous les mondes possibles dans lesquels le chien l'attrape est ce dont il a peur. Au premier abord, cette description des deux situations ne présente pas d'ambiguïté, bien que peu économique en mondes. Pourtant, dans un monde possible avec beaucoup de chats similaires, de chiens similaires et de toits similaires, quel chat est la contrepartie de notre chat ? Dans un de ces mondes possibles peuvent se trouver à la fois un chat qui lui ressemble sur un toit semblable au sien et un autre chat comme lui dans les mâchoires d'un chien ; ce monde appartient-il à la fois à l'état des choses désiré et à l'état des choses craint ?

Convaincu qu'il y a là un problème, Quine propose une solution, ou en tout cas un début de solution adapté aux cas d'attitudes égocentriques ou intéressées, c'est-à-dire aux cas où l'individu considère un état de lui-même, une situation qui l'implique. Le chat est un tel individu en attitude égocentrique. Considérons que, en quelque sorte, l'individu *se repère* dans les mondes possibles : il faut munir les mondes d'axes spatiaux et d'un axe du temps, et surtout chaque repère doit avoir une origine fixe[52] dans le corps ou organisme de la contrepartie de l'individu, de préférence au centre de gravité de la glande pinéale. Si un monde présente deux chats suffisamment similaires à notre chat, il y a alors deux possibles origines.

[52] Nous essayons ici d'être concis, mais Quine, rigoureux, donne de nombreuses précisions sur cette fixité de l'origine, la rotation des axes, etc.

Notons bien que l'origine est spatiale et aussi temporelle, ou que, pour le dire ainsi, le chat est à l'origine spatiale au temps 0, ce qui l'identifie comme le chat en attitude ; lorsqu'il se trouve éloigné de l'origine sur l'axe du temps, il peut éventuellement en être éloigné spatialement (s'il s'est déplacé). Nous avons donc maintenant des mondes possibles *centrés*, et une classe de tels mondes fait un état de choses *centré*.

Quine reste néanmoins assez flou sur cette notion de centre. Il emploie le mot pour désigner un point géométrique (l'origine du repère, le centre de gravité), mais il est clair que les termes « *egocentric* » et « *self-centered animal* » évoquent un sujet (si ce n'est un centre psychique qui s'accorderait peu avec le physicalisme). Posons la question : quelle phrase « éternelle », soit vraie soit fausse pour tout le monde et une fois pour toute, a pour signification une proposition centrée ? En effet, le pronom « je » est fortement appelé dans la formulation, qui du coup interdit l'éternité. « Je serai dans quelques instants sur le toit et non dans la gueule du roquet », dirait notre félin *s'il était capable de construire la phrase*. On se demande même si Quine n'a pas choisi un animal pour éviter l'objection. Une contrepartie du chat n'est finalement pas repérée par des coordonnées, c'est l'origine qui est repérée par la contrepartie : celle-ci est comme la partie centrale d'un univers dont chaque partie, y compris elle, est définie à partir d'elle ; là serait le sens complet d'« auto-centré ». Le philosophe n'est pas satisfait par sa solution parce qu'en voulant se débarrasser des ambiguïtés dans le choix d'une classe de mondes, il revient au problème du « je » et des indexicaux. Mais pour ses auditeurs et lecteurs, il ouvre la voie vers le traitement et notamment la probabilisation de nouvelles propositions logiques : celles-ci ou leurs avatars dans le langage naturel seront dits « temporels », « indexicaux », « auto-localisants ».

### 1.4. Réalisme modal et théorie des états relatifs

Sur un aspect précis, la théorie des mondes possibles divise les philosophes. La majorité d'entre eux rangent les mondes dans des régions logiques uniquement, pensent que seul notre monde est « réel », « concret », « actuel », « physique », « extérieur », pour ne citer que quelques qualificatifs qui ont chacun leurs imperfections, et que tous les autres mondes sont des abstractions, quelle que soit la qualité de l'aide qu'ils apportent aux scientifiques. Mais quelques chercheurs, dont Lewis, trouvent plus raisonnable de penser qu'il existe, très concrètement, une pluralité de mondes causalement indépendants, qui occupent chacun tout l'espace-temps, ou « leur » espace-temps ; chacun de ces univers, inclusif, totalisant mais pas total, aurait le même statut ontologique que notre monde[53]. Cette dernière thèse controversée, qui porte le nom de *réalisme modal*, suit un principe de rasoir d'Occam *qualitatif* : il ne faut pas multiplier les *types* d'entités au-delà de ce qui est nécessaire, il faudrait donc préférer une pluralité de mondes tous concrets à une pluralité de mondes abstraits s'ajoutant à un monde concret[54].

Lewis lui-même distinguant la « ramification des mondes » et leur « divergence »[55], il convient de faire une différence très nette entre cette théorie des contreparties à « structure divergente statique de mondes »[56] qu'est le réalisme modal, et la théorie des mondes multiples, ou théorie des états relatifs, qui est une interprétation de la mécanique quantique selon laquelle, pour le dire vite, l'univers, aussi appelé « multivers », prolifère, se

[53] Dans ses premières pages, *De la pluralité des mondes* explique cela peut-être mieux que dans de vieux articles sur le sujet ; la difficulté de ce célèbre ouvrage de Lewis est, selon nous, qu'il mêle parfois une défense du recours aux mondes possibles (qu'ils soient logiques ou « réels ») à une défense du réalisme modal.

[54] Schmitt (2012), p. 156.

[55] D. Lewis (2007), p. 315 *sqq.*

[56] Schmitt (2012), p. 162.

ramifie continuellement par émergence de structures, souvent appelées « mondes » ou « branches », de telle sorte que, lorsqu'une mesure quantique peut renvoyer différents résultats, tous ces résultats coexistent dans des structures.

Cette mention de la théorie des états relatifs n'est pas inutile. En effet, comme en témoignent encore aujourd'hui les récits sur les voyages vers des « mondes parallèles » dans la littérature et le cinéma de science-fiction, la théorie a été pendant longtemps très mal vulgarisée jusqu'à être confondue, justement, avec un réalisme modal lui-même dévoyé[57]. Après que son père, le physicien et mathématicien Hugh Everett, l'a proposée, encore imprécise, dans sa thèse de doctorat en 1956, elle fait travailler l'imagination des non-spécialistes. Elle est ignorée ou rejetée pendant des années par la communauté scientifique. Lewis en entend vraisemblablement parler dans les années 1970, alors qu'elle est défendue par quelques chercheurs[58]. Mais nous pensons que ce qu'il entend a des chances d'être proche d'une contrefaçon présentant les « mondes » comme des copies imparfaites de plus en plus nombreuses et dissemblables, une contrefaçon qui subit l'influence des réflexions philosophiques autour des mondes possibles et qui, en retour, représente un poids, une séduction qui attire le philosophe vers le réalisme modal. Entendons-nous : vers la fin du 20e siècle, le réalisme modal s'est développé dans la littérature en se souciant de cohérence et de dialogue avec ses nombreux adversaires logiciens et métaphysiciens, sans faire appel à la physique quantique. Lewis n'a vraiment étudié l'interprétation d'Everett qu'à la toute fin de sa

[57] Nous pourrions aussi évoquer la fascination pour les espaces n-dimensionnels, et la fameuse quatrième dimension. Ainsi que les voyages dans le temps.

[58] Elle est notamment popularisée par Bryce DeWitt en 1973 dans *The Many-Worlds Interpretation of Quantum Mechanics* (Princeton University Press).

vie et s'est montré plutôt critique[59]. Néanmoins, la possibilité qu'une théorie physique, même déformée, ait encouragé son positionnement réaliste ne peut pas être écartée.

Ce qui rend également inévitable la précédente brève évocation de la théorie des états relatifs, ce sont les publications de philosophes de la physique qui la lient aux problèmes d'auto-localisation. Notamment, une controverse a opposé *Peter* Lewis, David Papineau et Victor Durà-Vilà au sujet d'une question épineuse : peut-on défendre une solution tiériste au problème de la Belle au bois dormant tout en soutenant l'interprétation d'Everett, laquelle semble favoriser le demisme ? Nous en dirons un mot dans le quatrième chapitre. Une discussion plus récente entre Darren Bradley et Alastair Wilson[60] montre comment un débat sur les interprétations de la mécanique quantique peut se transformer en problème d'auto-localisation et mêler mondes possibles et univers multiples. En effet, « notre monde » (au sens métaphysique de l'expression) est possiblement un univers non proliférant, possiblement un multivers proliférant ; s'il est un univers, un observateur sur le point de mesurer un spin (disons le spin *x* d'un électron dans un état propre de spin *z*) n'a qu'un futur dans lequel il verra soit le résultat « spin *up* », soit le résultat « spin *down* » ; mais si le monde est un multivers, cet observateur se divise en une « partie quantique » qui voit *up*, et sa partie jumelle qui voit *down*. Ainsi, bien que l'observateur n'envisage que deux types possibles de mondes, il doit considérer que dans chaque monde-multivers il a plusieurs localités ou identités ; on voit l'analogie avec la Belle au bois dormant qui double ses positions temporelles dans les mondes où la pièce de monnaie tombe sur pile. Sur la base de cette réflexion, on pourrait construire un argument philosophique (évidemment discutable) en faveur de la théorie d'Everett,

[59] C'est ce que montre un des derniers documents écrits par Lewis, *How many Lives has Schrödinger's Cat?*, que l'on trouve dans Jackson et Priest (2004), p. 4-23.

[60] Bradley (2011a, 2015) et Wilson (2014).

de la même façon qu'une Belle tiériste, après un réveil dans l'expérience, élabore un raisonnement qui rend pile plus probable que face. Notre future analyse, dans ce chapitre, du problème de la machine à dupliquer montre la difficulté, pour un agent rationnel, de probabiliser des propositions localisantes après fission ou clonage.

### 1.5. Brandon Carter et le principe anthropique

Un autre événement attaché à l'idée des univers multiples et intéressant la cosmologie et la philosophie va, bien plus que la thèse everettienne, orienter la théorie de l'auto-localisation et les résolutions de ses paradoxes : c'est la formulation du principe anthropique par Brandon Carter lors d'un symposium de l'Union astronomique internationale en 1973. Le texte de l'intervention de Carter est publié l'année suivante sous le titre « *Large number coincidences and the anthropic principle in cosmology* ». Inspiré par plusieurs savants, de Dirac à Dicke, l'astrophysicien souhaite réagir à l'usage immodéré d'un « principe copernicien » selon lequel nous n'occupons pas une « position *centrale* privilégiée dans l'Univers ». Notre admiration pour l'intelligence humaine et la vie sur Terre, nos observations de la répartition de la matière dans l'Univers, notre théorisation de l'expansion de celui-ci, nos estimations de son âge, de ses constantes fondamentales, etc., nous invitent à remarquer et méditer une loi de bon sens. Celle-ci peut être énoncée faiblement ou fortement[61] :

– Principe anthropique faible : « Nous devons nous attendre à constater que nous nous trouvons à un endroit de l'Univers nécessairement

[61] Toutes ces citations sont tirées de Carter (1974).

privilégié au sens où il est compatible avec notre existence en tant qu'observateurs. »[62]

– Principe anthropique fort : « L'Univers (et donc les paramètres fondamentaux dont il dépend) doit être tel qu'il admet la création d'observateurs en son sein à un moment donné. Pour paraphraser Descartes, « *cogito ergo mundus talis est* ». »[63]

Ce qui ne fait guère de doute, c'est que ces « principes » prennent appui sur une tautologie[64] : puisque j'observe, ce que j'observe est une nature qui m'a permis d'être un observateur (si c'était autre chose, je ne serais pas là pour l'observer). Le principe faible est un remodelage de cette tautologie après identification de la « nature » à une portion très restreinte de l'Univers, un endroit privilégié qui permet l'apparition d'êtres vivants et d'êtres conscients, un îlot, peut-être non unique, quelque part dans un océan cosmique très souvent hostile. En ce qui nous concerne, c'est la surface de la Terre habitée notamment par nos congénères, les êtres humains. Interpréter le principe fort est plus délicat, d'autant plus que se joue son acceptation ou son rejet par les scientifiques. Dans un possible sens, il n'est autre que le principe faible mais avec identification de la « nature » à *notre* univers entier, qui serait un monde privilégié parmi une multitude d'univers étrangers, aux paramètres très différents, qui d'une façon ou d'une autre coexisteraient, ou bien se succéderaient, idée qui n'est plus un truisme mais une prise de position scientifique et philosophique difficile, d'où l'adjectif « fort » qui remplace « faible ». Le problème réside

[62] « […] we must be prepared to take account of the fact that our location in the Universe is *necessarily* privileged to the extent of being compatible with our existence as observers. »

[63] « […] the Universe (and hence the fundamental parameters on which it depends) must be such as to admit the creation of observers within it at some stage. To paraphrase Descartes, 'cogito ergo mundus talis est'. »

[64] Mosterín (2004).

dans les mots employés par Carter, notamment « doit être », « création » et la sentence latine d'inspiration cartésienne qui, sous une apparence tautologique, voudrait dire en réalité : je pense, donc l'Univers est « taillé pour la pensée », est créé pour une fin qui est la vie et la Raison. La saveur finaliste du texte de Carter étant sensible, on peut supposer que le principe est fort parce qu'il nous dit que les constantes fondamentales ne sont pas un hasard, qu'un dieu les a finement réglées. Toutefois, s'il n'apparait pas que l'auteur songe à un quelconque multivers au moment où il rédige *Large Number Coincidences*, d'autres de ses textes le suggèrent fortement. Carter (2006) juge que le principe anthropique évite deux écueils : le point de vue anthropocentré pré-copernicien et le nouveau point de vue cosmologique de l'univers homogène[65].

Les chercheurs ont donc interprété de diverses façons les principes faible et fort, et parfois ont apporté des modifications dommageables. Ainsi, les physiciens John Barrow et Frank Tipler, auteurs de *The Anthropic Cosmological Principle* à la fin des années 80, furent critiqués à la fois par des adversaires du raisonnement anthropique, tels que l'astrophysicien français Christian Magnan qui constate une dérive pseudo-scientifique et un retour de l'anthropocentrisme, et des philosophes adeptes tels que Nick Bostrom, qui considère que Barrow et Tipler n'ont pas compris Carter[66]. Bostrom estime par ailleurs que plus de trente « principes anthropiques » ont été formulés dans la littérature[67].

[65] Barrau (2007), p. 52-55, estime que le principe anthropique fort a été déformé par des penseurs finalistes en « principe anthropotéléologique », à l'opposé des intentions de Carter, dont le texte de 2006 et le regret d'avoir qualifié le principe d'« anthropique » confirment qu'il songeait à des modèles cosmologiques ou des univers aux constantes fondamentales différentes.

[66] Bostrom (2002), p. 47-50.

[67] *Ibid.*, p. 6.

Saisir le principe anthropique par la tautologie qui l'imprègne, c'est le comprendre un peu, mais manquer l'essentiel. Ce que ses premiers énoncés n'affirment pas encore assez clairement, mais que Carter et d'autres auteurs expliciteront à partir des années 1980, c'est que le *cogito* cartésien, reformulé pour l'occasion « Je suis un observateur », est contre toute attente une *information* de la plus haute importance, qui renseigne celui qui la saisit vraiment en tant que telle sur sa place dans l'univers, sur la qualité de son environnement, sur la quantité de ses voisins « humains », ou « observateurs » plus généralement. Le principe anthropique me signifie que, dès que je me découvre observateur, je détiens une preuve qui *a posteriori* rend extrêmement probable l'hypothèse que je me trouve au sein d'une nature féconde en êtres conscients, alors qu'une telle localité était *a priori* quasiment une impossibilité au vu de l'immensité, belle mais habituellement stérile, du cosmos. Selon Leslie (1987), le raisonnement anthropique a un grand potentiel et serait prédictif de nombreuses découvertes qui ne sont encore que des hypothèses négligées, sur les formes de vie extraterrestres par exemple. Le principe doit être nuancé et complété, mais encore aujourd'hui les philosophes ne s'accordent pas sur la manière, et le « biais anthropique », ou biais d'observation, nous menace quand nous tentons de répondre aux nouveaux défis scientifiques qui propulsent notre esprit vers les contrées lointaines et inexplorées de tout ce qui est ou *peut* être. Nous reviendrons immanquablement là-dessus lorsque nous rencontrerons quelques problèmes liés à l'auto-localisation.

Nous avons donc, cachée dans le principe anthropique, une règle qui modifie la *probabilité* d'une hypothèse après la prise en compte d'une information, grâce à un calcul implicite utilisant des probabilités conditionnelles selon toute vraisemblance. Nous reconnaitrons peut-être la conditionalisation *bayésienne*. Cette ultime interprétation de la pensée de Carter ne pouvait pas être connue par Lewis lorsqu'il rédigea son *Attitudes De Dicto and De Se*, bien qu'il soit un bayésien convaincu. Le rappel des

rudiments du bayésianisme est-il nécessaire ? La prochaine section commence justement par là…

## 2. Théorie des probabilités

La pensée de David Lewis est indéniablement marquée par quelques nouveautés dans la théorie des probabilités, notamment la diversification des approches bayésiennes, l'apparition du propensionnisme et les premières formulations du principe d'inférence directe.

### 2.1. Les principales règles du bayésianisme

Le bayésianisme est une interprétation épistémique des probabilités : elles sont entendues comme des intensités, des degrés de croyance, qui ne sont pas arbitraires, en tout cas pas purement arbitraires, mais se conforment à certaines normes de rationalité[68]. Selon la norme fondamentale qui est d'ailleurs la seule suivie par les bayésiens classiques, les degrés de croyance d'un agent rationnel obéissent aux lois du calcul des probabilités, autrement dit la mesure du crédit que l'agent accorde à un moment donné aux propositions qu'il n'est pas absurde de probabiliser (de « croire plus ou moins ») est une distribution qui satisfait les axiomes de Kolmogorov ou une axiomatique plus convaincante si elle existe. Ramsey dans les années 1920, puis de Finetti dans les années 1930 ont construit un argument pragmatique, l'argument du pari hollandais (*Dutch book*), pour justifier l'obéissance aux lois du calcul : si un agent a l'habitude de parier sur la base de ses degrés de croyance, et si ceux-ci violent le calcul des probabilités, on pourra toujours planifier et lui proposer une série de paris qu'il accepterait alors même qu'elle le conduirait à une perte certaine ; les

[68] Cozic et Drouet (2009) est la source principale, mais pas exclusive, de cette section.

bayésiens classiques ont démontré qu'un agent est invulnérable à ce type d'escroquerie si ses degrés de croyance obéissent au calcul[69]. D'autres arguments encore montrent que violer le calcul introduit de l'irrationalité dans l'action, mais ils sont aujourd'hui largement remis en cause, en particulier, justement, parce qu'ils négligent la dimension épistémologique des croyances en les réduisant à leur rôle dans l'action. Certaines tentatives pour résoudre les problèmes d'auto-localisation les plus coriaces utilisent des paris hollandais sophistiqués.

Le bayésianisme classique est subjectiviste puisque deux agents également rationnels et disposant des mêmes informations peuvent croire à des degrés différents s'ils ne sont contraints que par les axiomes des probabilités. Quelques chercheurs, dont Edwin Jaynes, ont proposé d'autres exigences de rationalité pour éliminer l'arbitrarité. L'ambition est carnapienne : ces auteurs considèrent la théorie des probabilités comme une extension de la logique. Une norme selon eux indispensable, qui fait du bayésianisme une approche empiriquement basée, est la conformité avec les données disponibles : la confiance qu'un agent doit accorder, au temps *t*, à une proposition A doit appartenir au plus petit sous-intervalle de [0 ; 1] qui contient toutes les valeurs compatibles avec les informations concernant A qu'il détient au temps *t*. Celles-ci sont principalement d'ordre statistique, ce sont des fréquences observées en prenant les classes de référence appropriées. Pour éliminer toute arbitrarité, au moins une norme de rationalité supplémentaire est nécessaire. La plus en vogue est la norme de neutralité, ou de maximisation de l'entropie : l'idée est qu'un degré de croyance doit être, parmi ceux compatibles avec les données disponibles, le plus éloigné des valeurs extrêmes ; mais isoler un degré n'est généralement pas possible, il faut considérer un groupe de croyances dont les degrés sont à choisir de façon à ce que l'ensemble maximise l'entropie. C'est une tâche

[69] Gillies (2000), p. 50-51, 59-61.

complexe mais nous n'avons pas besoin d'en parler plus longtemps. Disons seulement qu'il existe un seul ensemble de degrés de croyance qui respecte le calcul des probabilités, prend en compte les informations et maximise l'entropie[70]. Le bayésianisme qui prône ces trois exigences peut être dit objectiviste. Notons que les bayésiens objectivistes sont encore minoritaires au moment où Lewis produit ses grands textes sur la théorie des probabilités.

Comme son nom l'indique, le bayésianisme voit en Thomas Bayes un précurseur. Le pasteur britannique est l'auteur, au milieu du 18e siècle, du fameux théorème indispensable à la révision des probabilités dite *conditionalisation* qui occupe une place centrale, principielle dans la théorie des degrés rationnels de croyance, en tout cas jusqu'à la fin du 20e siècle. Regardons de plus près cette révision dans sa version doxastique. Les deux mesures du crédit qu'un agent rationnel accorde à ses croyances juste avant d'apprendre une information E (*a priori*) et juste après avoir appris E (*a posteriori*) sont notées respectivement $P(\cdot)$ et $P_E(\cdot)$. Ce sont des distributions de probabilité qui satisfont les axiomes de Kolmogorov. Supposons $P(E) > 0$. Alors l'agent qui apprend E doit maintenant attribuer à une hypothèse H la probabilité conditionnelle de H sachant E (probabilité de H si E) :

Conditionalisation : $P_E(H) = P(H \mid E)$

Grâce au théorème de Bayes, il est possible de donner ces deux formules qui faciliteront les calculs de la nouvelle probabilité de H (étant donné E) dans la plupart des applications de cette révision :

$$P_E(H) = P(E \mid H).P(H) / P(E)$$

$$= P(E \mid H).P(H) / (P(E \mid H).P(H) + P(E \mid \neg H).P(\neg H))$$

[70] Cozic et Drouet (2009).

Bien entendu, la probabilité de H reste inchangée si H et E sont indépendantes ; elle peut être diminuée ou augmentée sinon, et parfois jusqu'aux extrêmes 0 ou 1. On dit que H est confirmée par E si la probabilité de H étant donné E est supérieure à la probabilité initiale de H. Nous avons là la racine d'une théorie de la confirmation, et effectivement cette théorie, objet d'une vaste littérature, est une des grandes réalisations du bayésianisme[71].

### 2.2. Propensionnisme et inférence directe

Bien que les degrés de croyance soient privilégiés par Lewis dans tous ses raisonnements, il ne délaisse pas la réflexion sur la pertinence d'une probabilité ontique et s'intéresse à l'interprétation propensionniste, que Karl Popper expose notamment en 1959 dans l'article sobrement titré *The Propensity Interpretation of Probability*. En pensant aux problèmes de la mécanique quantique, Popper obtient la conviction qu'une probabilité, inobservable et difficilement mesurable, est pourtant « physiquement réelle » : c'est une propriété dispositionnelle d'une configuration expérimentale, une tendance à faire venir un état des choses donné, qui peut être quantifiée par un nombre de l'intervalle unité, approché par la fréquence de cet état des choses parmi tous les résultats d'expériences répétées dans les mêmes conditions, de très nombreuses fois, effectivement ou virtuellement. Cette *propension* est comparable à la force newtonienne davantage qu'à la potentialité aristotélicienne, car ce n'est pas une propriété inhérente à un dé, une pièce ou tout autre objet individuel, mais une propriété *relationnelle* d'un dispositif expérimental ou, si l'on veut simplifier, d'un arrangement de *plusieurs* choses ; de même, une force ne s'exerce sur un objet massif qu'en présence d'au moins un autre corps. Le

[71] Pour de nombreuses précisions concernant la conditionalisation, on se reportera à Gillies (2000), qui fait une distinction entre un bayésianisme logique (chap. 3) et un bayésianisme subjectiviste (chap. 4).

propensionnisme se désintéresse du problème de la série (in)finie d'événements de l'approche fréquentiste et donne un vrai sens à la probabilité d'un événement singulier, même s'il parait surtout adapté pour interpréter les probabilités de phénomènes quantiques tels que la production d'un photon par un atome radioactif dans un délai court. Malheureusement, il ne garantit pas que les mesures de propensions aient les propriétés formelles des probabilités ; il montre aussi de la fragilité dès qu'il s'interroge sur le sens de la probabilité conditionnelle[72].

La probabilité-propension va réformer la question nouvelle de l'alignement des probabilités ontiques et épistémiques. Depuis *A theory of probability* publié par Reichenbach en 1949, cette question n'envisageait comme probabilités ontiques que des fréquences relatives, elle était donc par exemple liée au difficile choix de la classe d'événements de référence ; mais elle était surtout liée au problème de l'inférence directe[73]. Explorons rapidement ce point.

Le problème de l'inférence directe peut être posé de cette façon : comment corriger le syllogisme statistique[74] grâce aux énoncés de probabilité, de telle façon que la probabilité (et non plus la vérité) d'un fait singulier soit logiquement amenée d'une statistique ou d'une fréquence relative donnée en prémisse ? Supposons que Pierre, 20 ans (en 2016), habite au Puy-en-Velay, et que 90 % des gens comme lui ont déjà goûté la lentille verte du Puy. L'inférence directe pourrait s'écrire : la probabilité statistique d'avoir déjà goûté la lentille verte du Puy quand, comme Pierre,

[72] Cozic et Drouet (2009).

[73] Thorn (2012).

[74] Le syllogisme statistique, ou syllogisme proportionnel, est un raisonnement inductif qui conclut le particulier à partir d'une généralisation vraie *dans la plupart des cas*. Il est souvent proposé sous la forme : La plupart des *X* sont *Y* ; or *a* est *X* ; donc *a* est *Y*. C'est un raisonnement fallacieux néanmoins très utilisé par le commun et étudié par les logiciens pour ses qualités inattendues.

on habite au Puy et qu'on a 20 ans est 0,9 ; donc la probabilité logique que Pierre ait déjà goûté la lentille verte est 0,9 (ou encore, si je suis bayésien, je dois croire au degré 0,9 que Pierre a déjà goûté la lentille verte). Mais la validité de l'inférence repose sur un principe d'alignement intuitif. Est-il possible de le préciser et de le formaliser ? David Miller propose en 1965, lors d'un colloque de philosophie des sciences[75], le principe qui porte son nom, parfois appelé « principe de l'inférence direct » lorsqu'il est précisément un pont entre probabilité *statistique* et probabilité *logique*, c'est-à-dire lorsqu'il manifeste les choix d'origine de son auteur, visiblement inspiré par Carnap. Téméraire et direct, Miller commence par faire l'impasse sur le problème des classes et de l'événement singulier et, prenant un exemple, appelle simplement probabilité statistique la probabilité d'obtenir 6 au prochain jet de dé, sous-entendant que ce lancer est le dernier d'une longue série. Il dit ensuite que si $\langle p(a) = r \rangle$, où $a$ est un événement, p une mesure de probabilité statistique et $r$ un réel de l'intervalle unité, est une information et si A est la proposition selon laquelle $a$ se produit, alors la probabilité logique de A, c'est-à-dire son degré de confirmation par $\langle p(a) = r \rangle$, est intuitivement $r$ :

Principe de Miller : $P(A, \langle p(a) = r \rangle) = r$

Deux raisons font que ce n'est qu'à partir de 1965 que les chercheurs se penchent sérieusement sur cette formule qui n'est pas révolutionnaire au premier abord et se trouvait sûrement dans bien des têtes avant cette date. La première est qu'elle est, paradoxalement, à la fois une nouvelle déclaration de scission entre une probabilité ontique et une probabilité épistémique, et l'affirmation de leur relation intime, sur laquelle Carnap était tout de même hésitant. La seconde est qu'elle conduit à des contradictions cachées, dont le « paradoxe de l'information » que Miller

[75] Miller (1966) reproduit les notes distribuées lors du colloque.

met en évidence[76]. Ce principe d'alignement est ainsi doublement paradoxal ; mais il n'est pas dit que ces paradoxes aient des natures si différentes et que leurs solutions ne se rejoignent pas. Quoi qu'il en soit, une possible aberration tapie derrière la simplicité formelle est crainte. Penser le principe prend du temps et Miller lui-même est persuadé qu'il n'est pas au bout de la recherche.

Jusqu'à aujourd'hui, une littérature peu intéressée par le propensionnisme va continuer à travailler sur l'inférence directe, à aménager l'ancienne version de son principe, à la charger de nouvelles protections pour écarter les problèmes qu'elle peut engendrer[77]. Mais plusieurs philosophes vont associer le travail de Miller à la théorie de la propension. Popper lui-même est un des premiers à commenter le paradoxe de l'information dans le numéro du *British Journal for the Philosophy of Science* où Miller le fait publier. En 1971, *The Matter of Chance*, regard perspicace sur les théories probabilistes, mentionne brièvement la controverse autour du principe de Miller dans le premier chapitre consacré au propensionnisme[78] : Hugh Mellor, auteur de cet ouvrage connu par David Lewis, rapproche de la propension la probabilité de l'événement singulier telle qu'elle est perçue par un certain fréquentisme, et prédit l'avenir serein d'un principe d'alignement corrigé. La réforme propensionniste du principe est engagée. Nous verrons bientôt que le

[76] Miller (1966) « démontre » que 1/4 = 1/6 grâce à une manipulation astucieuse de la formule.

[77] Par exemple, Paul Thorn a proposé récemment cette inférence directe évoluée : « If *A* is justified in believing that E[freq(T|R)] ∈ V and c ∈ R, then *A* has a defeasible reason to believe that PROB(c ∈ T) ∈ V, so long as E[freq(T|R)] ∈ V is *relevant* to the value of PROB(c ∈ T) for *A*. » Nous noterons plusieurs précautions : l'agent considère la *pertinence* de l'appartenance d'une fréquence *espérée* à un certain *ensemble de valeurs*, pour alors avoir une raison *défectible* d'estimer une probabilité. Thorn (2012) explique en détail ces notions.

[78] Mellor (1971), p. 67.

bayésien Lewis apporte une contribution *cruciale*, c'est-à-dire au *croisement* de ses réflexions sur l'auto-localisation et sur l'interprétation double des probabilités.

## 3. David Lewis, le sage analytique

Lewis est un touche-à-tout prolifique. En 1976 paraissent, certes, le premier texte dont nous allons étudier une section, à savoir *Probabilities of Conditionals and Conditional Probabilities*, mais aussi, par exemple, un article sur les paradoxes du voyage dans le temps, un autre sur l'identité personnelle, une notion que l'on sait chère à Arnold Zuboff, un des pères du paradoxe de la Belle au bois dormant. Quand on lit Lewis, on comprend que derrière son attachement à populariser, parfois non sans humour, les raisonnements avec les mondes possibles, il y a des interrogations profondes, celles de son maître Quine et d'autres encore. Avec lui, rien ne va de soi, son bayésianisme ne va pas de soi, le statut ontologique des mondes ne va pas de soi, tout comme ne vont pas de soi nos conceptions de la conscience, du temps et des notions les plus anciennes et les plus fuyantes. Nous n'exagérons pas si nous le rangeons parmi les philosophes originaux compris trop tard, ceux qui dérangent les conventions et nous invitent à envisager les choses d'une tout autre manière. Notre premier sentiment est qu'il cherche des solutions dans le fantastique et l'incroyable ; pourtant, de son point de vue, sa méthode consiste à accorder une grande importance au sens commun et aux croyances ordinaires[79]. Si les philosophes quittent le commun pour tendre vers le sage, Lewis, lui, n'est jamais vraiment parti, parce qu'il espère bien que le point de départ et celui d'arrivée sont intimes.

[79] Schmitt (2012), p. 156.

### 3.1. Le conditionnel de Stalnaker et l'imaging

*Probabilities of Conditionals and Conditional Probabilities*, qui est comme son nom l'indique un texte sur la relation entre probabilité conditionnelle et probabilité d'un énoncé conditionnel, est certes cohérent ; pourtant il ne cherche pas à être cohérent à tout prix, il explore plutôt l'étendue de la cohérence, va jusqu'à sa limite, et sa limite incertaine. Là, la cohérence consiste seulement à savoir qu'*aujourd'hui* le sens de nos constructions mentales nous échappe en partie, qu'*aujourd'hui* nous devons battre en retraite, et dans quelques années nous reviendrons plus forts, si ce n'est nous alors nos successeurs. Cette limite est touchée dans la section « *Probabilities of Stalnaker Conditionals* ». Elle est peut-être en lien étroit avec l'auto-localisation.

Lewis propose d'examiner le conditionnel de Robert Stalnaker, le philosophe qui selon lui a fourni le meilleur travail sur la thèse selon laquelle les probabilités conditionnelles sont des probabilités de conditionnels. Les conditions de vérité du connecteur > qui lie la proposition antécédente A et la conséquente C sont celles-ci : A > C est vrai si et seulement si la révision des faits la moins drastique qui rendrait A vraie rendrait aussi C vraie. Stalnaker conjecture que cette interprétation rendrait égales les probabilités P(A > C) et P(C | A) tant que P(A) n'est pas nulle[80].

Lewis estime que cette conjecture est fausse, mais pas totalement : si les deux probabilités ne sont pas égales en général, la probabilité du conditionnel a bien certaines des propriétés caractéristiques des probabilités conditionnelles. Il nous invite à raisonner avec les mondes possibles pour mieux le comprendre. Si une totalité de faits possible correspond à un monde possible, une révision correspond à une transition d'un monde à un autre. La révision de Stalnaker consiste à atteindre, à partir d'un monde *w*,

[80] Stalnaker traite aussi le cas P(A) = 0, mais il est inutile d'en parler ici.

le monde *le plus proche* pour lequel A est vraie, monde noté $w_A$. Ainsi $A > C$ est vrai dans $w$ si et seulement si C est vraie dans $w_A$. Maintenant, considérons une distribution de probabilité P sur les mondes possibles, supposés en nombre fini. Un agent apprend que A est vraie. Par conditionalisation, il devrait déplacer les probabilités des mondes-$\neg$A, les répartir sur les mondes-A de façon à protéger les probabilités relatives (à « conserver les proportions », le rapport des probabilités de chaque couple de mondes-A). Pourtant le conditionnel de Stalnaker ne correspond pas à cette révision classique, mais à une autre révision que Lewis appelle *imaging*. Par imaging, la probabilité d'un monde-$\neg$A est seulement déplacée vers le monde le plus proche où A est vraie. Donc la nouvelle distribution P′, dite image de P, est telle que :

$$\forall\, w' \in \mathrm{W},\ \mathrm{P}'(w') = \sum_{\{w \in \mathrm{W} : w_A = w'\}} \mathrm{P}(w)$$

Prenons un exemple. Vous téléphonez à un ami dont la passion est de fabriquer de petits objets en pierre ou en bois, jamais en d'autres matériaux. Il vous annonce qu'il est en train de sculpter une des deux pièces simples d'un couple mortier-pilon. Il y a donc quatre possibilités exclusives et conjointement exhaustives : il sculpte, soit un mortier de pierre (MP), soit un mortier de bois (MB), soit un pilon de pierre (PP), soit un pilon de bois (PB). Vous ne connaissez rien à la technique du pilonnage. Supposons que, suivant certaines raisons ou intuitions glanées à l'écoute de votre ami, vous attribuiez *a priori* une très faible probabilité à MP et une relativement bonne (entre 1/4 et 1/3) aux autres possibilités. Votre ami vous apprend maintenant qu'il ne sculpte jamais de mortier en bois. Vous annulez donc la probabilité de MB mais devez augmenter les trois autres ou certaines d'entre elles pour qu'elles totalisent 1. Comment allez-vous vous y prendre ? Si à ce moment précis la conditionalisation est pour vous la seule révision rationnelle, vous augmenterez le poids des trois possibilités restantes en n'accordant qu'un extrêmement faible supplément à l'improbable MP. Par contre, s'il vous est plus naturel de penser que les

mondes- ¬MB les plus proches des mondes-MB sont des mondes-MP, voire des mondes-PB, en tout cas pas des mondes-PP, trop différents, alors vous n'allez peut-être pas modifier la probabilité de PP, vous allez préférer augmenter celle de MP. En d'autres termes, si selon vous, dans ce cas précis, fonder une probabilité *a posteriori* sur la probabilité d'un conditionnel de Stalnaker a plus de sens que la fonder sur une probabilité conditionnelle classique, si vous accordez à ¬MB > MP sens, attention et forte probabilité, c'est une autre règle de réajustement doxastique que vous vous préparez à appliquer : la règle de l'imaging.

Ce que veut dire Lewis, c'est que la probabilité du conditionnel de Stalnaker fait en quelque sorte office de probabilité conditionnelle dans une révision concevable, si ce n'est rationnellement applicable, qui est semblable dans la forme à la conditionalisation, mais qui aboutit souvent à d'autres résultats. Après le gain d'une information E, la nouvelle probabilité d'une hypothèse H sera :

– par conditionalisation : $P_E(H) = P(H \mid E)$

– par imaging : $P_E(H) = P(H > E)$

Ces possibles révisions des probabilités sont toutes deux minimales, mais pas dans le même sens. La première, comme nous l'avons dit, ne distord pas le « profil » des groupes de probabilités, c'est-à-dire conserve les ratios, les égalités, les inégalités. La seconde n'autorise aucun déplacement gratuit de probabilité de mondes devenus impossibles vers des mondes dissemblables.

Quelque chose doit nous étonner à ce stade. Quand Lewis rédige cet article, ainsi que sa suite, *Probabilities of Conditionals and Conditional Probabilities II*, qui parait dix ans plus tard, son but est de défendre l'épistémologie bayésienne orthodoxe en relevant le défi des avocats de la thèse selon laquelle les probabilités conditionnelles sont des probabilités de

conditionnels. Il réussira finalement à démontrer qu'il ne peut pas exister un connecteur, → ou >, tel que P(C | A) = P(A → C). Cependant, comme on le voit ici, les connecteurs, principalement celui de Stalnaker, lui donnent matière à réflexion, si bien qu'il conçoit une révision doxastique originale qu'il place en face de la conditionalisation, en concurrence, mais en ne manifestant ni optimisme ni pessimisme vis-à-vis de son avenir. Ce n'est pas l'attitude d'un bayésien orthodoxe, mais celle d'un parieur audacieux et patient, qui sur une intuition mise gros en vue d'un possible gain très lointain, qui connait les risques et attend un coup de pouce du destin, pas immédiat mais différé.

### 3.2. La réception de l'imaging par les philosophes

À quoi peut bien servir un imaging ? Quelle situation, quelle condition, quel détail le rendrait soudain plus rationnel que l'inférence bayésienne traditionnelle ? Dubitatifs sont les lecteurs de Lewis. L'imaging n'est pas pris au sérieux dans un premier temps. On montre assez facilement qu'il n'a pas la propriété intuitive de « préservation » qu'a la conditionalisation ; cette propriété fait qu'une proposition crue pleinement, avec certitude, est toujours tenue pour certaine après le gain d'une information compatible avec les croyances initiales[81]. Penchons-nous une dernière fois sur l'exemple du mortier-pilon : imaginons que nous tenons pour certain, avant d'apprendre que la possibilité mortier-en-bois doit être écartée, qu'aucun mortier en pierre n'est en train d'être taillé par l'ami passionné ; nous comprenons qu'après le gain d'information un imaging rigoureux ferait passer de 0 à un nombre strictement positif la probabilité de mortier-en-pierre. C'est très étranger à nos habitudes, profondément suspect. Seule une information incompatible avec les croyances initiales,

[81] Cf. Gärdenfors (1988) pour des précisions sur les propriétés de la conditionalisation et la trahison de l'imaging.

donc une information elle-même étrangère et suspecte, a, lorsqu'elle est acceptée malgré tout, un pouvoir si infirmant qu'il déconcerte, déchire des certitudes.

Quelques publications vont pourtant trouver un intérêt à l'imaging à partir des années 1990, notamment Katsuno et Mendelzon (1992) et Walliser et Zwirn (2002). Les croyances pleines, ainsi que les croyances partielles, peuvent être modifiées pour deux raisons : ou bien nous apprenons quelque chose sur notre environnement, et dans ce cas nous *révisons* nos croyances en conséquence ; ou bien nous apprenons que notre environnement a changé, et dans ce cas nous *mettons à jour* nos croyances. Mettre à jour peut transformer une certitude en incertitude, car il ne s'agit plus de « mieux » croire *ce qui est*, comme si notre monde était dans un état stable, figé dans le temps, mais de corriger certaines croyances parce que le monde *en devenir* est passé d'un état à un autre. Les règles de conditionalisation peuvent être dérivées de la transcription probabiliste des postulats de la *révision* des croyances pleines ; mais ce que l'on a surtout découvert, c'est que les règles de l'imaging peuvent être dérivées d'une transcription probabiliste de postulats un peu différents, « optimisés » pour la *mise à jour* des croyances pleines. Nous n'entrerons pas dans les détails car nous ne croyons pas que Lewis a en tête une telle idée. Le plus simple des exemples suffira pour l'instant. Je suis certain qu'il y a un et un seul objet dans cette enveloppe, soit un gros billet de banque (probabilité 1/2), soit un vieux tract publicitaire (probabilité 1/2). J'apprends plus tard qu'un voleur qui prend l'argent et seulement l'argent vient de regarder dans l'enveloppe. Je dois alors croire que l'enveloppe est vide (probabilité 1/2) ou bien contient un tract (probabilité 1/2). Que s'est-il passé ? Suivant une règle indiscernable de celle de l'imaging, j'ai déplacé la probabilité 1/2 de la classe de mondes-billet sur *la plus proche* classe de mondes-sans-billet, qui est évidemment la classe des mondes-enveloppe-vide qui avait auparavant une probabilité nulle. Dans ce contexte de mise à jour due à une

modification physique du système de l'enveloppe et de son contenu, l'imaging fonctionne bien, il brise assez naturellement ma certitude qu'il y a un objet dans l'enveloppe, il n'augmente pas la probabilité des mondes-tract, trop dissemblables.

Des philosophes ont donc trouvé un formidable intérêt à l'imaging. Mais notre souci à nous, c'est l'intuition de Lewis, lequel ne semble pas comprendre son nouveau concept comme une mise à jour des croyances partielles, en tout cas se préoccupe davantage de la proximité des mondes ou des états de choses que de leur devenir. S'il n'est pas sévère avec sa découverte peu orthodoxe, c'est probablement qu'il la voit s'appliquer à des regroupements de mondes particuliers que notre exemple du mortier-pilon a ratés. Et trois ans après *Probabilities of Conditionals*… il rédige son *Attitudes De Dicto and De Se*. Nous pensons aux mondes centrés. La tentation est grande, mais n'est-elle pas irrationnelle ?

Dans leurs très récentes études des problèmes d'auto-localisation, quelques philosophes ont suggéré que la conditionalisation classique est souvent inadaptée lorsque les possibles sont des mondes centrés. Chris Meacham, fin analyste des difficultés des textes lewisiens, jette les bases d'une « conditionalisation compartimentée » : les mouvements de probabilité après disqualification d'une possibilité centrée affecteraient en priorité, s'il y en a, des possibilités centrées voisines, c'est-à-dire des centres dans le même monde possible. Mieux encore : à la même période, Mikaël Cozic rapproche l'imaging, par l'intermédiaire des travaux de Walliser et Zwirn, de la révision (peut-être mise à jour) des possibilités centrées qu'il trouve la plus intuitive, et qui s'avère être celle de Meacham[82]. Relier imaging et auto-localisation n'est peut-être pas si absurde. En 1976, Lewis ne parle pas encore de mondes centrés, en tout cas

[82] Cf. Cozic (2007) et Meacham (2008). Un manuscrit de Meacham circulait depuis 2003. Le point de vue « double demiste » de Cozic et Meacham sur la Belle au bois dormant sera abordé au chapitre 4.

pas dans une publication, mais il connait la réflexion de Quine sur la difficulté de trouver la contrepartie de l'agent en attitude dans les mondes possibles présentant plusieurs sujets qui ressemblent à l'agent, et il participe en outre, rappelons-le, à un débat sur l'identité personnelle et la continuité mentale dans la survie (à un lavage de cerveau, à ma fission en deux jumeaux, à ma fusion avec une autre personne, etc.). Deux mondes possibles *m* et *w* sont dits proches lorsque le passage de *m* à *w* conserve plus qu'il ne transforme les faits de *m* ; or, la transformation qui permet de passer d'un monde centré sur un sujet au même monde centré sur un autre sujet similaire est en quelque sorte une simple translation de sujet à sujet, qui conserve la totalité des faits : la proximité serait donc excellente dans ce cas. Il est très possible que Lewis ait en tête cette excellence lorsqu'il juge concevable et pertinent l'imaging. Pour illustrer et éclaircir ce que nous venons de dire, analysons un problème d'auto-localisation évoquant les préoccupations de Lewis à l'époque, au lieu de donner les exemples moins appropriés de Meacham et de Cozic ; mettons donc de côté les positions temporelles et expliquons plutôt comment un agent rationnel devrait réviser la probabilité épistémique de ses possibles *identités*.

### 3.3. Le problème de la machine à dupliquer

Des scientifiques qui ne se posent pas trop de questions éthiques sont parvenus à fabriquer une machine à dupliquer un être humain en quelques minutes : le double est (juste après sa production) une copie physique quasi parfaite de l'original, c'est en outre un être conscient aussi rationnel que l'original, ayant les souvenirs, les humeurs, les goûts de l'original, si bien qu'il serait persuadé d'être l'original s'il n'avait pas connaissance de la duplication. Les savants disposent de deux cobayes adultes en bonne santé : Alice et Bruno. Ils leur annoncent qu'ils vont les endormir et les dupliquer l'un après l'autre : il y aura donc après ce processus quatre sujets qui seront séparés, placés dans quatre chambres éloignées, et qui n'auront à leur réveil

aucun moyen de savoir s'ils sont des originaux ou des copies. Alice et Bruno savent que les savants, fous ou pas, ne mentent pas.

Les duplications ont bien lieu. Voici que Bruno se réveille dans sa chambre : il a l'impression d'être Bruno mais, aussi bizarre que cela puisse paraitre, il a maintenant des raisons de croire qu'il est peut-être un double de Bruno avec la mémoire de ce Bruno original. Supposons qu'il croit être l'original au degré 1/2 et la copie au même degré. Un des scientifiques entre dans la pièce et lui demande : « À quel degré dois-tu croire que Bruno a été dupliqué en premier, avant Alice ? » Bruno doit répondre 1/2 s'il n'a pas de raison de croire cette hypothèse (appelons-la B) avec une intensité différente de celle de la croyance contraire. Il peut maintenant envisager ces quatre possibilités exclusives et conjointement exhaustives et attribuer à chacune la probabilité 1/4 :

AO : « Alice a été dupliquée en premier et je suis original »

AC : « Alice a été dupliquée en premier et je suis une copie »

BO : « Bruno a été dupliqué en premier et je suis original »

BC : « Bruno a été dupliqué en premier et je suis une copie »

Le savant poursuit : « Je t'ai posé cette question parce qu'il se trouve que la copie issue de la première des deux duplications vient de décéder dans son sommeil. Cause inconnue. Les autres sujets sont réveillés et se portent très bien. » Bruno doit-il toujours croire au degré 1/2 qu'on l'a dupliqué en premier ? Beaucoup de spécialistes des problèmes d'auto-localisation répondraient par la négative : Bruno ne devrait maintenant croire cette hypothèse B qu'au degré 1/3. D'autres chercheurs n'en seront pas aussi sûrs. Résumons les arguments que pourrait avancer chaque camp, et voyons où se loge l'imaging.

Une donnée importante est révélée : trois sujets et non quatre ont survécu, plus exactement le double de la première personne dupliquée est décédé. Du point de vue de celui qui donne l'information, celle-ci ne sort pas de l'ordinaire, elle n'est que le rapport d'un fait, elle est « éternelle » ; du point de vue du cobaye laissé dans l'incertitude, c'est autre chose. En effet, BC n'est plus une possibilité pour Bruno (qui sait évidemment qu'il n'est pas un cadavre), autrement dit il vient d'apprendre $\neg$BC = AO $\vee$ AC $\vee$ BO. De son point de vue, l'information est auto-localisante, elle exige qu'il se localise à présent dans la peau du Bruno d'origine si un monde dans lequel Bruno est le premier dupliqué est actuel, dans la peau de l'un ou de l'autre des deux « jumeaux » si un monde dans lequel Bruno est le second dupliqué est actuel. Si l'on suit la règle de conditionalisation privilégiée par le bayésianisme, on peut calculer par exemple la probabilité *a posteriori* de BO :

$$\begin{aligned} P_{\neg BC}(BO) &= P(BO \mid \neg BC) \\ &= P(\neg BC \mid BO).P(BO) / P(\neg BC) \\ &= 1.(1/4) / (1 - 1/4) \\ &= 1/3. \end{aligned}$$

De même, $P_{\neg BC}(AO) = P_{\neg BC}(AC) = 1/3$. Sans surprise, la conditionalisation répartit équitablement l'ancienne probabilité 1/4 de BC sur les trois possibilités restantes afin de conserver leur équiprobabilité. Pour Bruno, croire B revient maintenant à croire BO (BC étant une impossibilité) ; par conséquent il ne peut plus répondre 1/2 mais bien 1/3 à la question du savant. Notons aussi qu'il accorde à présent deux fois plus de crédit à la possibilité qu'il soit Bruno original (AO $\vee$ BO) qu'à la possibilité qu'il soit une copie (AC). Observons enfin que l'information auto-localisante $\neg$BC modifie par conditionalisation des croyances auto-localisantes telles que BO *mais aussi des croyances éternelles* telles que B.

Conditionalisation sur ¬BC

| | | | | |
|---|---|---|---|---|
| AO : **1/4** | | BO : **1/4** | AO : **1/3** | BO : **1/3** |
| | ↖ | ↑ | | |
| AC : **1/4** | ← | BC : **1/4** | AC : **1/3** | BC : **0** |
| Probabilités *a priori* | | | Probabilités *a posteriori* | |

Un bayésien plus critique ne peut pas être satisfait par cette démonstration en raison d'éléments peu clairs. Intuitivement, l'élimination de BC, certes, augmente la probabilité de BO. Oui, des mouvements de probabilités ont lieu. Mais la probabilité de B doit-elle chuter jusqu'à 1/3 ?

Supposons que la démonstration précédente, utilisant la conditionalisation, résume assez bien le raisonnement d'un parangon de rationalité. Supposons qu'Alice et Bruno sont tous deux des parangons de rationalité qui ont un ensemble de croyances initiales identiques (nous ne parlons que des croyances liées à leur expérience de duplication), qui ont reçu les mêmes informations depuis leur réveil, donc qui sont au courant du décès de la première des deux copies d'êtres humains et qui savent qu'il y a quelque part dans ce monde une Alice et un Bruno réveillés en bonne santé. Alors Bruno doit croire B au degré 1/3 quand Alice doit croire B… au degré 2/3. En effet, elle raisonne avec des hypothèses centrées sur elle (le « je » de l'hypothèse BO, par exemple, la désigne) et elle traduit en ¬AC et non en ¬BC la donnée selon laquelle elle ne peut pas être la copie décédée ; donc elle croit ¬B au degré 1/3, et B au degré 2/3. Allons plus loin : si Bruno est bien un modèle de rationalité, s'il sait qu'Alice l'est aussi, il peut donc imaginer qu'en ce moment même Alice originale apprend ¬AC et en conséquence attribue à B une probabilité différente de celle qu'*il* attribue à B. Éternelle, B n'a pas plusieurs significations en fonction du sujet qui la pense ; en outre B est crue par chaque sujet à un degré conclu au bout d'une démarche logique éloignée de toute arbitrarité, malgré la nécessité de peser au passage des croyances auto-localisantes.

Aussi B semblait la dernière croyance sur laquelle Bruno pouvait s'attendre à être en désaccord avec Alice. Nous avons du mal à comprendre comment un agent rationnel peut soutenir le degré 1/3 en sachant qu'il serait 2/3 chez un autre. C'est ce qui nous pousse à croire que notre hypothèse de départ (la démonstration qui utilise la conditionalisation résumerait assez bien le raisonnement d'un parangon de rationalité) est fausse.

La réflexion qui précède n'est pas assez convaincante ? Certes. Poursuivons. Avant d'apprendre le sort tragique du premier double, Bruno, indifférent, rendait équiprobables les quatre possibilités AO, AC, BO et BC, en sachant que plein de futures données pourraient modifier cet équilibre d'une façon ou d'une autre. Mais tout de même, apprendre la mort dans son sommeil d'un double, dont la vie physique fut donc éphémère et la vie consciente nulle, c'est recevoir une information assez particulière. Imaginons que le double de Bruno, copié avant Alice, décède dans son sommeil : il ne peut évidemment pas entendre un scientifique annoncer cette mort et l'identifier à la sienne, autrement dit il n'a jamais l'occasion de savoir BC avec certitude et d'annuler les probabilités de AO, AC et BO. Quand Bruno original apprend la mort du premier des deux doubles, il peut éliminer BC parce qu'il fait le constat de son état conscient, le constat qu'il n'est pas le cadavre annoncé. Mais quel autre constat pouvait-il faire ? Sûrement pas celui de la cessation de son existence. En outre il s'est réveillé avec l'impression d'être l'original, mais ne sait pas qu'il l'est ; par manque d'information il croit aussi B et ¬B au même degré. Et voilà que, juste parce qu'il étend prudemment les possibilités de son identité à deux individus de ce monde plutôt qu'à un seul, l'annonce de la possible mort du double (à moins que ce soit celui d'Alice) déséquilibrerait les crédits qu'il accorde à ¬B et B, des propositions éternelles ? En fait, tout se passe comme si l'information donnée par le scientifique manquait d'efficacité ou n'était pertinente qu'à moitié, ne

permettait que des mouvements de probabilités minimaux, parcimonieux. Et c'est là qu'on retrouve l'imaging.

L'inférence bayésienne qui, par habitude, fait correspondre à AO, AC, BO et BC quatre classes de mondes possibles puis, après élimination d'une classe, partage sa probabilité sur tous les mondes restants, semble manquer un détail important : BO et BC ont une proximité essentielle du fait qu'elles peuvent être interprétées comme possiblement vraies dans le même monde mais pas pour le même sujet. C'est dans un même monde que Bruno coexiste avec son double, et lorsqu'il pense BO, lorsqu'il a une attitude-BO, il se localise non seulement dans un monde où Bruno est dupliqué avant Alice, mais aussi dans la peau de Bruno original plutôt que dans celle de la copie : au sens de Quine il se repère dans un monde centré. Que la vérité ou la crédibilité de BO dépendent davantage de la vérité ou de la crédibilité de BC que de la vérité ou crédibilité de AC, que l'élimination de BC modifie prioritairement le crédit de BO, ce sont là des pensées qui passeront difficilement pour irrationnelles. Supposons qu'après l'annonce du scientifique, Bruno annule la probabilité de BC, n'augmente que celle de BO, qui passe de 1/4 à 1/2, et laisse ainsi inchangée la probabilité de B. Comment le lui reprocher et lui montrer un manque de rationalité ? Il faudrait par exemple trouver un système de paris acceptables proposés avant et après la révision des croyances, peut-être même avant et après la duplication, qui conduirait Bruno à une perte certaine au bout du compte. La tâche est très malaisée ; un système efficace serait *peut-être* obtenu en donnant un portefeuille commun à Bruno et à sa copie et en engageant dans des paris Bruno et sa copie si elle est vivante, mais qu'il n'y ait pas assurément un seul parieur rendrait contestable la conclusion de l'irrationalité de l'acceptation des paris.

Imaging sur ¬BC

| | | | |
|---|---|---|---|
| AO : **1/4** | BO : **1/4** | AO : **1/4** | BO : **1/2** |
| | ↑ | | |
| AC : **1/4** | BC : **1/4** | AC : **1/4** | BC : **0** |
| Probabilités *a priori* | | Probabilités *a posteriori* | |

### 3.4. Croyances *de se* et mondes centrés

Dans son argumentation sur les probabilités de conditionnels, Lewis a jugé bon de mettre côte à côte, en concurrence, conditionalisation et imaging, comme nous l'avons fait dans cette analyse du problème de la machine à dupliquer. Bien que nous ne sachions pas ce qu'il avait précisément en tête, il fut conduit à publier *Attitudes De Dicto and De Se* trois ans plus tard, puis *Subjectivist's Guide to Objective Chance* l'année d'après, deux textes qui sont comme des remèdes à deux besoins, à deux envies qui nous tourmentent durant l'analyse des mésaventures de Bruno et de son amie Alice : tout d'abord, évidemment, le désir d'en apprendre davantage sur l'auto-localisation (qu'est-ce qu'un centre exactement ? l'opposition entre monde centré et monde non centré doit-elle être nuancée ?) ; et puis le besoin de l'autre versant des probabilités. Quand son interprétation épistémique a du mal à arbitrer le litige des révisions sur information, le bayésien cherche un indice physique objectif.

*Attitudes De Dicto and De Se* présente l'auto-localisation dans une forme propre à Lewis. L'idée principale est qu'une attitude est une localisation d'un sujet dans les mondes et/ou dans le temps, donc l'auto-assignation d'une *propriété* ou de plusieurs propriétés partagées par un groupe d'individus (éventuellement altermondains ou au contraire colocataires d'un même monde), ce qui signifie que toute croyance *de dicto*, c'est-à-dire éternelle au sens de Frege ou propositionnelle au sens de Quine, se place sous le concept plus général de croyance *de se*. Lewis est

strict sur la définition d'une proposition : c'est une classe de mondes possibles, et seulement ensuite quelque chose comme une phrase. Considérer que les attitudes n'ont pour objets que des propositions, c'est se limiter de façon absurde, c'est ne pas voir que tout individu appartient à une population qui s'étend à la fois dans les mondes et l'espace-temps. Le réalisme modal, qui tente d'expliquer que les mondes sont habités par des êtres très concrets, pèse évidemment de tout son poids dans cette perspective lewisienne, malgré l'emploi dans certains cas de l'expression atténuante et conciliante « espace logique » :

> Nous sommes dispersés non seulement dans l'espace logique, mais aussi dans le temps et l'espace ordinaires. Nous pouvons avoir des croyances par lesquelles nous nous localisons dans l'espace logique. Pourquoi pas aussi des croyances par lesquelles nous nous localisons dans le temps et l'espace ordinaires ? Nous pouvons nous attribuer des propriétés de sorte qu'elles correspondent à des propositions. Pourquoi pas aussi des propriétés qui ne correspondent pas à des propositions ? Nous pouvons nous identifier comme membres de sous-populations dont les limites suivent les bords des mondes. Pourquoi pas aussi comme membres de sous-populations dont les limites ne suivent pas les bords des mondes ? Pourquoi pas ? Il n'y a pas de raison ! […] Ces croyances sont des attitudes dont les objets devraient être considérés comme des propriétés auto-assignées plutôt que comme des propositions tenues pour vraies.[83]

Ainsi, l'originalité de Lewis relativement à d'autres théoriciens de l'auto-localisation consiste à refuser une séparation entre croyances *de se* (auto-localisantes) d'un côté et croyances *de dicto* (éternelles) de l'autre. « Je suis né en 1975 » situe l'agent en attitude dans des mondes proches du nôtre (où, au moins, naitre en 1975 a un sens) et surtout parmi une sous-population de gens nés en 1975, c'est une croyance *de se* mais pas *de dicto* puisque, indexicale, elle est vraie ou elle est fausse selon l'agent qui la soutient. En revanche, « Le verre est un isolant électrique » *situe* le croyant dans des mondes où le verre a cette propriété isolante, c'est une croyance

[83] D. Lewis (1979), p. 519.

*de dicto*, autrement dit une croyance *de se* qui identifie l'agent comme membre d'une sous-population de *contreparties*, qui ne le situe pas possiblement dans une sous-population habitant le même monde. Ainsi « la *de se* subsume la *de dicto*, mais pas *vice versa*. »[84]

Arrive une réflexion de grande importance, puisque nous l'assortirons aisément avec des problèmes de type Belle au bois dormant :

> Jusqu'à présent, j'ai considéré les sujets des attitudes comme des gens ordinaires, ou à peu près. Et les gens sont des continuités, étendues dans le temps. Mais certains cas de croyances *de se* peuvent être mieux compris si nous considérons le croyant non comme une continuité mais plus ou moins comme une tranche temporelle, momentanée de celui-ci. Auparavant j'ai supposé que chaque sujet d'attitudes habitait seulement un monde, même si, comme le pensent quelques-uns, les personnes sont étendues à travers les mondes. Je fais maintenant une supposition parallèle avec l'extension dans le temps.[85]

Lewis veut dire que chacun d'entre nous a tendance à se confondre avec ses contreparties dans les mondes possibles mais reconnait qu'il est bien une partie isolée qui n'habite qu'un monde, et que certains cas d'auto-localisation nous invitent à construire une analogie avec notre extension dans le temps et à nous considérer comme une partie temporelle de notre être continu. Examinons le cas de l'insomniaque qui ne sait plus quelle heure il est. Il sait à peu près dans quelle sorte de monde il vit et quelle place il occupe dans l'espace-temps, il sait qui il est, quel jour on est, ce qu'il fait. La connaissance qui lui manque n'est pas propositionnelle, et si elle est une assignation de propriétés, ce n'est pas à son être continu. Ce qu'il se demande, c'est quelle *partie* temporelle de lui-même il est. Quand il cherche quelle heure il est, il essaie de se situer dans une sous-population de l'ensemble de ses « moments ». Plus tard, Lewis appellera ces parties

[84] *Ibid.*, p. 521.

[85] *Ibid.*, p. 527.

« *individuals-at-times* »[86]. Pourtant, attaché au vocabulaire de Quine, il ne cessera de les appeler « centres ».

Le centre de Quine est un individu repéré dans un monde et qui est possiblement l'agent en attitude. Le centre de Lewis est cela aussi, avec la dimension du temps mise davantage en évidence. C'est une partie temporelle. Seul un couple ($s$, $t$) où $s$ est un sujet et $t$ un instant ou un intervalle de temps peut être appelé centre par Lewis ou d'autres philosophes, même critiques, tels que Stalnaker. Un monde centré est un couple ($m$, $c$) où $m$ est un monde et $c$ un centre. Une possibilité centrée correspond à une classe de mondes centrés ; une telle classe *peut* contenir un nombre $n_1$ de couples $(m_1, \cdot)$ et un nombre $n_2 \neq n_1$ de couples $(m_2, \cdot)$.

La théorie de la décision est-elle modifiée si nous considérons que toutes les attitudes sont *de se* ?

> Réponse : très peu. Nous remplaçons l'espace des mondes par l'espace des mondes centrés, ou par l'espace de tous les habitants des mondes. Tout le reste est comme avant. Quels que puissent être les points de l'espace des possibilités, nous avons des distributions de probabilité sur l'espace et des attributions de valeurs d'utilité aux points. […] Mais puisque l'espace des possibilités n'est plus l'espace des mondes, ses régions auxquelles sont associés degrés de croyance et de désirabilité ne sont plus des propositions. Ce sont des propriétés.[87]

Cette réponse est très commentée par les théoriciens de l'auto-localisation. Elle est tellement imprécise qu'on peut l'utiliser pour apporter un soutien à des argumentations très différentes. Ce que nous croyons est que Lewis ne parle pas de révision des probabilités. Aussi, tout est ouvert ; notamment rien n'indique qu'il renonce à l'imaging ou qu'il ne reconnait

[86] D. Lewis (2001) utilise ce vocabulaire.

[87] D. Lewis (1979), p. 534.

pas une conséquente proximité des centres colocalisés dans le même monde possible[88].

### 3.5. L'eccéitisme et sa critique

*Attitudes De Dicto and De Se* répond à des objections qualifiées d'« eccéitistes ». L'eccéitisme selon Lewis transforme comme par impossible un moment, un lieu ou une identité en un état des choses, ramène toute attitude *de se* à une attitude *de dicto* équivalente, toute propriété ou localité à une proposition, une classe de mondes, et déclare *éventuellement* que les phrases indexicales équivoques du langage ordinaire ont donné l'impression que les attitudes avaient des objets autres que les propositions. Analysons brièvement trois situations à la manière de philosophes eccéitistes.

J'emprunte un petit chemin de campagne. Arrivé à une bifurcation, je peux prendre à gauche ou à droite. Je sais qu'un des deux chemins me fera marcher pendant deux heures, tandis que l'autre, un raccourci, me fera marcher pendant une heure trente, pour arriver à la même destination. J'ignore où est le raccourci, mais comme j'ai toujours pris à gauche dans des situations similaires, je choisis encore de prendre à gauche : il est alors midi. Au bout d'une marche constante, j'arrive à destination. Si je pense que j'ai suivi le raccourci, j'en déduis qu'il est treize heures trente ; inversement, si je crois qu'il est treize heures trente, je crois aussi que j'ai choisi le raccourci, autrement dit je crois que je vis dans un monde où le sentier de gauche est le raccourci, et non dans un monde où le sentier de droite est le raccourci. Dans ce contexte, « il est treize heures trente » est équivalente à « le sentier de gauche est le raccourci ». Pourtant, seule la dernière proposition, *de dicto*, est clairement proposition au sens d'expression d'une classe de mondes.

[88] D. Lewis (2001) a justement besoin de cette notion de « colocalisation ».

Il nous est tous arrivé que la mémoire nous joue un tour et nous fasse perdre un repère dans le cours de notre semaine de travail. Je ne sais plus si on est jeudi ou bien vendredi. Un « monde où on est jeudi » n'a pas de sens, jeudi est un moment et non un état du monde. Et pourtant, quand je veux par exemple peser « on est jeudi » relativement à « on est vendredi », je m'efforce de me rappeler des faits passés, de déduire leurs effets possibles, et d'estimer quelque chose comme une proportion, parmi tous les mondes possibles, de mondes où une contrepartie de moi-même, dans la semaine en cours, se demande *jeudi* quel jour on est. Quand je soutiens, pleinement ou à un certain degré, la croyance *de se* « on est jeudi », il me semble que je crois en réalité « le monde que j'habite est un monde où c'est jeudi qu'il m'arrive de me demander quel jour on est ». Cette dernière croyance est *de dicto*, malgré la présence des pronoms de la première personne, puisque ceux-ci désignent des contreparties possibles et non plusieurs colocataires possibles (d'un même monde).

Pour me rendre à l'IHPST, je m'enfonce dans les couloirs du métro parisien près de la gare d'Austerlitz, puis je monte dans un train de la ligne 10. Arrivé à la station Cluny-La Sorbonne, je repense à un aphorisme de Lao-tseu que j'ai lu la veille. Quand je sors de ma méditation, je m'aperçois que le train est sur le point d'atteindre une autre station, mais sur le moment j'ignore laquelle car, même si j'ai l'impression que mon esprit n'est pas resté absorbé longtemps, la durée réelle m'est inconnue ; il est donc possible que cette gare ne soit pas celle qui côtoie Cluny, Odéon, mais bien la suivante, où je dois m'arrêter, Mabillon ; il est même possible que ce soit une station encore plus éloignée, ce qui signifierait que j'ai raté mon arrêt. Furtivement, beaucoup d'indices me viennent à l'esprit : mon état de santé général, la consistance de ma réflexion taoïste, la distance estimée entre les stations, la fluidité de la ligne 10, la quantité, la disposition, le mouvement, le bruit de choses et de personnes dans le train, mon souvenir d'expériences similaires… La conjonction de plusieurs éléments effectifs

ou possibles me donne une certaine configuration possible du monde qui fait que c'est peu avant l'arrivée à Odéon que je délaisse mon intériorité et redevient présent au monde ; une autre configuration possible me ferait émerger près de Mabillon, etc. Un calcul approximatif me conduit par exemple à une certaine probabilité qu'en ce monde l'état des choses soit tel que je devais sortir de ma méditation près de Mabillon. Là encore, je ramène une croyance qui me localise dans l'espace-temps (« J'arrive à Mabillon ») à une croyance qui ne me « localise » que dans des mondes. Cette histoire lewisienne de propriétés et de centres est suspecte, donner trop d'importance aux propositions indexicales nous rend peut-être moins raisonnables, Frege nous avait dit de nous méfier…

Lewis répond : certes, quand nous ne savons plus en quel lieu, en quel temps nous sommes, ou encore qui nous sommes, c'est dans certains cas parce que nous souffrons d'un manque de connaissance sur le monde. Mais il n'est pas juste de contester l'authenticité d'une attitude *de se* non propositionnelle. Nous pouvons connaitre suffisamment le monde que nous habitons et alors croire ou désirer une propriété sans jamais nous situer plus amplement dans l'espace logique, sans jamais envisager une pluralité de mondes. C'est la différence entre le saint-croisé et le snob[89], qui pourtant tous deux cherchent une meilleure situation. Le chevalier croisé veut vivre dans un monde sans souffrance, donc faire partie d'une certaine population altermondaine, tandis que le snob veut vivre dans un meilleur quartier de la ville, de *cette* ville dans *ce* monde auquel il appartient. Le croisé veut être dans une meilleure partie de l'espace logique quand le snob veut être dans une meilleure partie de l'espace ordinaire.

Lewis expose le cas extrême des deux dieux[90], dont il tire une analyse qui selon lui réfute l'argument eccéitiste. Un monde est habité par deux

[89] D. Lewis (1979), p. 531.

[90] *Ibid.*, p. 520-524.

dieux : ils savent dans quel monde précis ils vivent, ils savent toutes les propositions vraies dans leur monde. Ils sont donc omniscients si l'on ne considère que des connaissances propositionnelles. Et pourtant, aucun d'eux ne sait lequel des deux il est. L'un vit au sommet de la plus haute montagne, l'autre au sommet de la plus froide montagne, et ils le savent, mais quand ils pensent « je vis sur la plus haute montagne », ils ne savent pas s'ils tombent juste. Si l'un des dieux apprenait que c'est lui-même qui vit sur la plus haute montagne, il aurait une connaissance supplémentaire, mais pas propositionnelle. Un eccéitiste répondrait, comme à son habitude, qu'un dieu ne peut pas être omniscient s'il ignore où il vit dans son monde, car cette ignorance est hésitation entre deux mondes possibles : un monde comme celui précédemment décrit, et un monde identique mais dans lequel les dieux ont échangé leurs places. Pour un dieu, savoir où il vit dans son monde revient à savoir qu'il vit dans un des deux mondes et pas dans l'autre, et c'est donc acquérir une connaissance propositionnelle. Lewis dit : non, ça ne change rien, l'eccéitiste fait à nouveau la confusion entre savoir que tel dieu vit sur la plus haute montagne et savoir que *soi-même* vit sur la plus haute montagne. En apprenant qu'il habite tel monde et non l'autre, le dieu complète son savoir propositionnel, certes, mais ne sait toujours pas si lui-même a la propriété qu'il rapporterait indexicalement : « Je vis sur la plus haute montagne ». La subtilité est là : le dieu qui connait toute proposition connait aussi la proposition qu'il exprimerait par cette phrase, et pourtant il ignore *quelle* proposition est exprimée par cette phrase.

Lewis peut donc conclure que les philosophes doivent adapter leurs discours à des attitudes *de se* non réductibles à des *de dicto*, et que son *Attitudes De Dicto and De Se* représente un essai réussi d'adaptation.

### 3.6. Le crédit sur la chance

1979 est une année riche pour Lewis qui publie des réflexions originales et parfois audacieuses dans plusieurs revues et démontre un certain goût pour les paradoxes philosophiques (notamment dans « *Prisoners' Dilemma is a Newcomb Problem* »). Comme en 1976, son bayésianisme subjectiviste est mis à rude épreuve, et dans ses analyses le manque de probabilités ontiques se fait sentir. Le professeur de Princeton cède enfin, et en 1980 parait dans un ouvrage collectif le célèbre *Subjectivist's Guide to Objective Chance*, qui présente un des principes d'alignement ontique-épistémique les plus connus, une loi qui joue notamment un grand rôle dans l'analyse des problèmes d'auto-localisation. D'ailleurs, Christopher Meacham, acteur majeur de la controverse *Sleeping Beauty*, fut amené en 2010 à produire un formidable article apportant des éclaircissements bienvenus sur le *Subjectivist's Guide* : « *Two Mistakes Regarding The Principal Principle* ». Notre commentaire doit beaucoup à ce texte de Meacham.

Lewis veut montrer une association subtile des deux faces de la probabilité et ainsi exalter un projet apparemment paradoxal d'interprétation « subjectiviste » des énoncés de probabilité ontique. Il commence par reconnaitre le sens et l'utilité d'une probabilité ontique et objective, différente de la probabilité épistémique subjective qui habite généralement ses textes. Il vante un dualisme du concept mais il le modère, il a conscience qu'il se distingue de Carnap sur des points importants :

> Nous, subjectivistes, concevons la probabilité comme la mesure d'une croyance partielle raisonnable. Mais nous n'avons pas besoin de faire la guerre aux autres conceptions de la probabilité, en déclarant que là où s'arrête le crédit subjectif commence le non-sens. Outre le crédit subjectif, nous devons aussi croire en la chance objective. La pratique et l'analyse de la science requièrent les deux concepts. […]

> Carnap a bien fait de distinguer deux concepts de probabilité, en insistant sur le fait que les deux sont légitimes et utiles et qu'aucun n'est fautif parce qu'il n'est pas l'autre. Pourtant je ne pense pas que Carnap ait choisi les deux meilleurs concepts. À la place de son « degré de confirmation », je mettrais le *crédit* ou *degré de croyance* ; à la place de sa « fréquence relative sur le long terme », je mettrais la *chance* ou *propension*, étant entendu qu'elle fait sens dans le cas de l'événement singulier. La division du travail entre les deux concepts sera un peu modifiée par ces remplacements. […]
>
> Avec les deux genres de probabilité, crédit et chance, nous pouvons avoir des probabilités hybrides de probabilités. (Pas des « probabilités de second ordre », qui suggèrent une sorte de probabilité auto-appliquée.) La chance du crédit ne nous retiendra pas. […] Le crédit sur la chance est plus important. Pour qui croit en la chance, celle-ci est un sujet propre à être cru. Les propositions sur la chance bénéficieront de degrés de croyance divers, et conditionnellement à elles d'autres propositions seront crues à des degrés divers.[91]

Ne nous trompons pas : c'est bien l'opposition ontique-épistémique qui est essentielle. L'opposition objectif-subjectif, qui n'est pas chez Carnap, est chez Lewis fragile : elle persiste tant que les deux concepts de probabilité s'insèrent dans des discours séparés, mais elle a tendance à s'estomper quand ils se lient dans un principe d'alignement, tel que le principe de l'inférence direct que le philosophe entend corriger. En effet, les deux concepts sont choisis et préparés pour être associés dans un même discours et dans une même loi. Notamment, le crédit, probabilité épistémique, et la propension, probabilité ontique, sont des fonctions ayant peut-être plusieurs arguments mais, bien souvent, un même argument objet : un événement singulier ou la proposition qui le rapporte. Autrement dit, la propension est préférée à la fréquence en vue d'un alignement avec le crédit, facilité par l'argument commun. Lewis avait apparemment en tête une certaine alliance des deux concepts avant même d'avoir une idée précise du concept ontique. Il décrit d'ailleurs cette alliance par

[91] D. Lewis (1980), p. 263-264.

l'expression « probabilités hybrides ». Le « crédit sur la chance » est une des formes (avec la « chance du crédit », laissée de côté) de la probabilité hybride qui mêle les deux concepts, il correspond bien à la manière d'un bayésien subjectiviste de parler de la chance dans un discours mixte.

Ce crédit sur la chance apparait d'abord comme crédit *au sujet de* la chance, comme le degré de la croyance d'un agent en une proposition de type X = « La propension à l'instant *t* de ce système aléatoire simple à faire venir le résultat *a* est *x*. » Pourquoi *t* ? Parce que pour Lewis une propension est nécessairement temporelle : avant la venue d'une des issues possibles, la chance de l'issue *a* appartient à l'intervalle unité, elle est vraiment *propension* ; après, elle est extrême (0 ou 1), autrement dit, soit *a* n'est pas réalisé, soit *a* est réalisé. Le monde est plus ou moins dans l'attente d'un événement futur daté ; passé cette date, l'événement est inscrit dans le monde à jamais. Ultimement, lorsque X est crue au degré 1, le crédit sur la chance se révèle crédit *aligné sur* la chance, c'est-à-dire que la proposition A = « Le résultat *a* est obtenu » doit être crue au degré *x*… *mais à une condition* : l'agent ne doit pas avoir appris d'information qui rend la connaissance de X de nul effet sur la croyance en A, ce que Lewis appelle une information *inadmissible*. C'est une différence majeure avec le principe de Miller.

L'admissibilité lewisienne n'est pas une notion parfaitement claire, y compris, en 2010, pour Chris Meacham. Le *Subjectivist's Guide* la relie à la notion de résilience d'une probabilité épistémique ou d'une incertitude. On lance une pièce de monnaie à midi. Vous êtes *sûr* que la pièce a, sur ce lancer au moins, une propension à tomber sur face de 1/2. Si vous n'avez pas d'autre information pertinente, vous devez croire au degré 1/2 que cette pièce lancée à midi tombe sur face. Si vous apprenez que la pièce a un défaut de fabrication, qu'elle a été lancée cent fois auparavant et que quatre-vingt-six faces ont été obtenus, et d'autres informations de ce genre admissibles à midi, susceptibles d'influencer votre croyance en face mais

seulement en influençant vos croyances sur les chances objectives de face, et si malgré tout vous restez certain que ces chances à midi sont 1/2, alors vous devez toujours croire en face au degré 1/2 ; cette incertitude de face étant fondée sur la certitude des chances de face, elle est stable, résiliente face à l'arrivée de nombreuses informations qui ne sont déstabilisantes qu'en apparence. Si vous obtenez plus tard certaines preuves disponibles seulement après le lancer, par exemple vous découvrez par vous-même que la pièce est tombée sur face, ou bien quelqu'un de confiance vous l'annonce, ou encore vous apprenez que le lancer de midi fait partie de dix lancers récents qui se sont conclus par neuf faces et un seul pile, vous devez maintenant croire en face à un degré proche de 1 : la résilience a atteint sa limite, vous ne pouvez plus aligner le crédit sur les chances objectives à midi. Des informations comme « la pièce lancée à midi tombe sur face » ou « la pièce lancée à midi fait partie de dix lancers qui se concluent par neuf faces » n'étaient pas admissibles *à midi*, même si elles sont compatibles avec la donnée de la propension 1/2 à midi. De telles informations viennent à bout de la résilience du crédit 1/2 initialement accordé à face, crédit qui ici augmente nettement.

### 3.7. Le principe principal et ses variantes

Nous sommes maintenant prêts à formuler ce principe « subjectiviste » qui « semble capturer tout ce que nous savons sur la chance »[92], fait entrer celle-ci autant que possible dans les raisonnements qui régissent notre espace probabilisé de mondes (éventuellement centrés), et fait oublier la différence, pourtant toujours essentielle, des deux concepts de probabilité grâce à une focalisation sur leur alliance dans une même tâche. Soient C la fonction de « crédit initial »[93] d'un agent rationnel, p une

[92] *Ibid.*, p. 266.

[93] C'est ce qu'en théorie de la décision certains chercheurs anglophones appellent aujourd'hui « *ur-priors* ».

fonction de chance temporelle dont l'argument objet est propositionnel, A une proposition du domaine de p (qui par exemple affirme la réalisation d'une certaine éventualité), *t* un instant, *x* un réel de l'intervalle unité, E une proposition compatible avec $\langle p_t(A) = x\rangle$ et admissible à *t*, donc qui ne peut pas influencer les croyances de l'agent à propos de A.

Principe principal : $C(A \mid \langle p_t(A) = x\rangle \wedge E) = x$

Meacham fait cette remarque : si nous supposons comme Lewis qu'il est rationnel d'être bayésien, alors nous sommes contraints par $P_I(\cdot) = C(\cdot \mid I)$, où $P_I$ est la fonction de crédit d'un agent dont l'information totale (*total evidence*) est I. Cela permet de remodeler le principe :

$P_{\langle p_t(A)=x\rangle \wedge E}(A) = x$

*Si nous précisons* que $\langle p_t(A) = x\rangle \wedge E$ est l'information totale de l'agent, une autre formule équivalente, utilisant la fonction de crédit actuelle P, est :

$P(A \mid \langle p_t(A) = x\rangle \wedge E) = x$

Dans l'article ultérieur « *Humean Supervenience Debugged* », Lewis change C en P imprudemment, ce qui est selon Meacham dommageable. Que Lewis reste vague sur la notion d'admissibilité est également regrettable. Meacham songe à d'éventuels mondes où le temps est une boucle, un peu comme dans les modèles de relativité générale de Gödel : dans de tels univers où un même événement appartient à la fois au passé et au futur d'un agent, il n'est pas facile de dire ce qui compte comme une information inadmissible, qui apparait comme une « information sur le futur » dans notre récent exemple de la pièce de monnaie. Mais sans aller aussi loin que Meacham, nous pressentons qu'interpréter l'admissibilité de telle ou telle façon influence l'analyse de cas où, comme dans la Belle au bois dormant, un sujet peut subir une amnésie qui efface complètement de sa mémoire une courte période de son histoire récente, et ainsi se retrouve

incapable de se repérer dans le temps aussi précisément qu'auparavant. Il est alors susceptible de récupérer les informations auto-localisantes perdues, et celles-ci sont peut-être pertinentes pour apprécier les causes, les natures ou les modalités possibles de l'amnésie. Selon que l'agent les jugera admissibles ou inadmissibles, il s'autorisera ou pas à avoir, à propos d'une possibilité, des croyances qui s'écartent des chances objectives. Nous aurons l'occasion de rencontrer une telle analyse lors de notre examen du problème de la Belle au bois dormant.

Le *Subjectivist's Guide* donne tout de même un autre indice : est certes admissible une proposition « historique » rapportant des faits survenus avant *t*, mais aussi une proposition sur la chance *elle-même*, ce qui expliquerait le sentiment déjà évoqué d'une mainmise du principe principal sur la chance. Ces considérations inspirent une tout autre écriture du principe, qui met de côté ou plutôt maquille la proposition admissible E. L'argument de base de p n'est plus le temps mais une proposition géante, la conjonction d'une théorie complète de la chance T, et de H, l'histoire complète jusqu'à l'instant *t* d'un monde où cette théorie est vraie. On a alors :

Principe principal reformulé : $p_{H \wedge T}(A) = C(A \mid H \wedge T)$

Lewis considère que cette version est à peu près équivalente à la première. Meacham démontre que les formules sont équivalentes à la condition qu'on adopte une définition précise de l'admissibilité, où T et H entrent en jeu[94] ; il ne dit pas que cette définition doit prévaloir. Notons surtout chez Lewis la volonté de placer la probabilité épistémique et la probabilité ontique de part et d'autre du signe égal. Est-ce un détail si la

[94] Meacham (2010) propose : « *E is admissible relative to* $\langle ch_t(A) = x \rangle$ *iff* $\langle ch_t(A) = x \rangle E$ can be expressed as a disjunction of some subset of the *TH*'s associated with $\langle ch_t(A) = x \rangle$. » La lecture de l'entière section 4 de l'article est nécessaire pour comprendre tous les détails.

probabilité ontique se situe à gauche ? Lorsque Meacham reproduit la formule avec ses conventions, il échange les deux termes, comme pour signifier que c'est plus naturel, que c'est le crédit qui s'aligne sur la chance et non l'inverse. D'autres commentateurs font de même.

Le principe principal a inspiré la littérature, qui a proposé plusieurs principes modifiés légèrement ou généralisés, mais aucun n'a eu le succès de l'original[95]. Même Lewis, associé à Ned Hall et Michael Thau[96], a souhaité corriger son propre travail en 1994, en constatant qu'il n'est pas cohérent avec ce qu'il appelle la « survenance humienne », une particularisation de la structure du monde qui se définit en gros par la présence de relations uniquement spatio-temporelles et de propriétés uniquement locales (représentant des caractéristiques intrinsèques de points spatio-temporels). Les développements difficiles de Lewis, Hall et Thau sont certes intéressants mais il serait vain de les exposer : non seulement leur « *New Principle* » manque d'élégance[97], mais en plus nous pouvons le considérer comme un cas particulier du principe de 1980, qui ne changerait pas notre étude des problèmes d'auto-localisation.

Lewis, bayésien mis en difficulté par ses propres thèses audacieuses, répond brillamment à l'appel insistant des théories ontologiques de la probabilité, en suggérant : « Je ne suis pas une victime mais un guide, ce n'est pas moi qui ai besoin des probabilités objectives mais elles qui ont besoin d'un subjectiviste comme moi. » Le principe principal n'est pas une règle pour objectiver tous les degrés de croyance d'un individu, comme si celui-ci était contraint en permanence par les probabilités lues dans le monde. C'est une règle pour croire à un degré *initial* la réalisation d'un événement particulier dont on connait la chance objective. Ce degré a une

[95] Pettigrew (2012), par exemple, compare plusieurs « principes principaux ».

[96] Les trois philosophes ont rédigé trois articles, publiés en 1994 dans la revue *Mind*.

[97] Nous le disons en accord avec Strevens (1995).

résilience qui peut ensuite céder face à l'arrivée d'informations « vraiment pertinentes ». Car il y a une pertinence apparente et des pièges à éviter, comme le montre l'exemple de la pièce de monnaie. Apprendre à midi qu'une pièce est tombée sur face lors des vingt derniers lancers est suffisant pour faire chanceler le crédit 1/2 accordé au résultat pile d'un imminent vingt-et-unième lancer lorsque ce crédit est simplement déduit du principe d'indifférence, mais n'est pas suffisant si le crédit est fondé sur la certitude d'une propension à midi. C'est la force de l'alignement, la force de la rencontre des deux visages de la probabilité. Ce qu'il nous faudra examiner de près, ce sont les paradoxes de l'auto-localisation où une mystérieuse force contraire est à l'œuvre, fâchant ceux qui veulent les résoudre.

## 4. L'auto-localisation jusqu'à aujourd'hui

Les théoriciens de l'auto-localisation sont nombreux. Se détachent néanmoins trois philosophes qui ont exposé des vues significativement différentes de celles de Lewis, ont complété sa théorie ou bien l'ont mêlée au raisonnement anthropique : John Perry a produit des textes importants à la fin des années 1970, Nick Bostrom autour de l'an 2000, et enfin Robert Stalnaker en 2008.

### 4.1. Perry et le problème de l'indexical essentiel

Le philosophe américain John Perry a travaillé sur les croyances auto-localisantes parallèlement à l'auteur d'*Attitudes De Dicto and De Se*, qui connaissait et a utilisé son premier article sur le sujet, « *Frege on Demonstratives* », paru en 1977, mais pas l'article plus mûr paru en 1979, « *The Problem of the Essential Indexical* ». Contrairement à Lewis, Perry ne s'intéresse pas à ce moment-là à la théorie de la décision ni à la théorie des probabilités, et il part directement du texte de Frege sans s'appuyer sur

celui de Quine. En revanche, l'identité personnelle est pour lui un objet de recherche important ; il semble d'ailleurs que ce soit le point commun de nombre de chercheurs qui ont perçu le pouvoir des indexicaux.

Selon Perry, Frege a commis une erreur : il a confondu la pensée et le sens et a fait de ce contenu appelé pensée ou proposition l'objet propre de la croyance. Mais une croyance a deux objets : le contenu et le *rôle* qui est la manière dont le contenu est cru et la cause d'une réaction du croyant. Lorsqu'à onze heures j'allume la télévision pour regarder mon émission favorite, ce n'est pas parce que je crois seulement que l'émission commence à onze heures, mais parce que je crois que l'émission commence à onze heures et qu'il est onze heures, autrement dit parce que je crois qu'elle commence *maintenant*. Dans différents contextes, la phrase « L'émission commence maintenant » renvoie à différentes propositions qui peuvent être vraies ou fausses, mais ne voir en l'indexical qu'un appel à un complément d'information, c'est rater le rôle qu'il joue dans la croyance. Dans un supermarché, Perry suit une trainée de sucre en poudre pour retrouver la personne qui, probablement, parcourt les allées avec un sac de sucre percé. Un employé vient à sa rencontre et lui dit en riant : « Vous semez le désordre ! » Perry s'aperçoit alors qu'il suivait les traces laissées par le sac qu'il transportait ! Quand l'employé croit le contenu < John Perry, semer le désordre, *t* >, il le croit avec le rôle de « *Vous* semez le désordre » et en conséquence de sa croyance il prévient le client. Celui-ci croit alors le même contenu mais avec le rôle de « *Je* sème le désordre », et va boucher le trou du sac. Pour expliquer leurs actions, les sujets usent d'indexicaux ; ceux-ci sont souvent « essentiels », ce qui veut dire chez Perry que leur remplacement par d'autres termes « détruit la force de l'explication »[98].

[98] Perry (1993), p. 35.

Toute croyance n'est pas localisation de soi pour Perry, loin de là. Une croyance auto-localisante est la croyance d'un agent sur où, quand ou qui il est dans le monde. Repérer un indexical essentiel peut aider à les différencier des croyances ordinaires, mais essentiel ne veut pas dire clairement énoncé. « On est le 2 août 2015 » exprime une croyance auto-localisante en masquant un peu l'indexical essentiel « aujourd'hui (est le 2 août 2015) ». « Il pleut aujourd'hui » présente en revanche un indexical non essentiel : la phrase est prononcée par un individu qui veut seulement dire quelque chose sur le ou son monde, dire que la pluie tombe à l'endroit où il s'exprime, le jour où il s'exprime, endroit et jour sur lesquels il ne s'interroge pas. « Il pleut aujourd'hui » n'exprime donc pas une croyance auto-localisante. Lewis aurait lui aussi jugé cette dernière croyance *de dicto*, mais il aurait ajouté que toute croyance *de dicto* est *de se*, puisqu'elle localise le croyant, certes dans l'espace logique uniquement, mais *le localise* quand même. Perry, on le voit, n'envisage surtout pas une telle subsumation.

Cette limitation du domaine des croyances auto-localisantes plait à certains philosophes qui les voient comme des objets inhabituels appelant de nouvelles règles de dynamique doxastique. Le souci, peut-être, ce sont les tabous qui tombent quand s'achève le 20$^{e}$ siècle. Pour Lewis, « proposition centrée » est certes un non-sens puisqu'une proposition est une classe de mondes, et pourtant l'expression va assez facilement faire son apparition chez quelques chercheurs, y compris chez Bostrom, Meacham ou Schwarz, des lecteurs et commentateurs de Lewis ; éventuellement, cela fera juste dire à celui-ci : « Attendez, vous ne semblez pas avoir ma définition d'une proposition… » Mais l'expression « proposition auto-localisante » va aussi être lue dans des publications parfois davantage inspirées par Perry ; un problème apparait alors : qu'est-ce que la conjonction d'une proposition auto-localisante (au sens de Perry) et d'une proposition éternelle ? Qu'est-ce que nous exprimons par « On est le 2 août

et il pleut » et quelle sera la dynamique des croyances ?[99] Finalement, il est difficile de préférer le compte de Lewis à celui de Perry, ou inversement. On voit que la question de la subsumation n'est pas rien.

### 4.2. Bostrom et le problème de l'auto-sélection

Le philosophe suédois Nick Bostrom est un spécialiste du raisonnement anthropique, qu'il met parfois en lien avec ses autres thèmes de prédilection (le transhumanisme, le clonage, le téléchargement de l'esprit…). Il prolonge certaines discussions auxquelles le philosophe canadien John Leslie a amplement participé dans les années 1980 et 1990, notamment à propos de l'argument de l'Apocalypse, principal paradoxe lié aux principes anthropiques, énoncé pour la première fois par Brandon Carter en 1983. L'argument de l'Apocalypse (*Doomsday argument*) est un raisonnement apparemment valide menant à une conclusion que la plupart des analystes trouvent irrecevable, à savoir la forte probabilité d'une fin (très) proche de l'humanité. En gros, de la même façon que chacun d'entre nous peut penser qu'il avait une très forte probabilité de naitre à une époque où la population du monde est de quelques milliards d'êtres humains plutôt qu'à une époque où elle n'était par exemple que de quelques millions, chacun peut aussi penser qu'il y a une très faible probabilité que son rang de naissance soit inférieur au millième du nombre total d'êtres humains qui ont vécu ou vivront sur Terre ou ailleurs dans l'Univers, c'est-à-dire une très faible probabilité que le nombre total d'êtres humains de toutes époques soit énorme, cent mille milliards par exemple ; à contrario, il y a une très forte probabilité que la fin de l'humanité soit proche, de telle sorte par exemple que seuls cent milliards d'individus au total la composent.

[99] Bartha (2006), notamment, pose ces questions au sein d'une analyse de la Belle au bois dormant. Dans cet article, les propositions éternelles sont appelées « substantielles ».

Bostrom s'interroge sur la nature de cette « probabilité » du raisonnement anthropique, tantôt ontique et objective, tantôt épistémique et susceptible de varier d'un sujet à l'autre, interprétée tantôt comme proportion, dans la population, d'individus ayant une propriété donnée, tantôt comme degré de croyance. Il remarque que Leslie est imprudent dans ce domaine[100], puis se concentre sur des travaux de Lewis. Il donne en 1999, à Nottingham, une conférence qu'il intitule « *A Subjectivist Theory of Objective Probability* »[101] en hommage au célèbre texte de Lewis, et prépare une dissertation doctorale qui doit associer une théorie de ce qu'il appelle alors l'« auto-sélection », c'est-à-dire la manière dont un observateur doit se considérer dans une population lorsqu'il prend conscience de sa qualité d'observateur, et une théorie de la chance. Pourtant, la dissertation finale sacrifie pratiquement le deuxième aspect au profit du premier.

Bostrom apprend l'existence d'un tout nouveau problème : la Belle au bois dormant. Il acquiert alors la conviction que les problèmes de l'auto-sélection et de l'auto-localisation appartiennent à un même genre et ne pourront être résolus que grâce à une théorie des « effets de sélection dans l'observation » et à la recherche de « biais » dans les raisonnements. L'argument de l'Apocalypse et la Belle au bois dormant ont des structures différentes mais pas en tous points. Dans les deux cas, il est demandé de probabiliser deux états de choses possibles : un état organisé de telle manière qu'il est (était, sera) observé peu de fois, et un état qui est (était, sera) observé un plus grand nombre de fois. Dans le *Doomsday*, les observations sont faites par des personnes différentes : une possibilité prévoit cent milliards d'observateurs, une autre cent mille milliards. Dans

[100] Bostrom (2000) critique la « chance relative à l'observateur » que Leslie croyait percevoir.

[101] Peut-être par erreur, Bostrom (2002) donne un autre titre, plus proche encore de celui de Lewis : « *A Subjectivist Theory of Objective Chance* ».

*Beauty* les observations sont faites à différents moments par un unique agent : une possibilité prévoit une observation (un réveil), une autre deux observations (deux réveils). Bostrom appelle *observer-moment* un bref segment de temps d'un observateur. Il parlera plus tard de « partie temporelle d'un agent » pour identifier son concept au concept lewisien de centre.

L'ouvrage *Anthropic Bias*, publié en 2002 et devenu ensuite une référence pour les spécialistes du raisonnement anthropique, est un essai de théorisation des effets de sélection dans l'observation. Bostrom explique notamment qu'un observateur *devrait* se sélectionner par « auto-échantillonnage ». Tout raisonnement anthropique doit faire l'hypothèse SSA (*Self-Sampling Assumption*) :

SSA : Un observateur devrait raisonner comme s'il était un échantillon aléatoire de l'ensemble de tous les observateurs dans sa classe de référence.[102]

Le refus du mot « principe » et l'emploi du conditionnel ne sont pas nécessaires, ils rappellent simplement que SSA peut être ou ne pas être complétée par une hypothèse anthropique controversée nommée SIA (*Self-Indication Assumption*) :

SIA : Étant donné le fait que vous existez, vous devriez (toutes choses égales par ailleurs) favoriser les hypothèses selon lesquelles beaucoup d'observateurs existent par rapport aux hypothèses selon lesquelles peu d'observateurs existent.[103]

Sont donc finalement concevables deux hypothèses anthropiques rivales : la première rejette SIA, l'autre l'intègre. Leurs formulations pourraient être celles-ci :

[102] Bostrom (2002), p. 57.

[103] *Ibid.*, p. 66.

SSA (rejetant SIA) : Toutes choses égales par ailleurs, un observateur devrait raisonner comme s'il était un échantillon aléatoire de l'ensemble de tous les observateurs dans sa classe de référence et colocalisés (c'est-à-dire séparés de lui dans l'espace-temps, jamais dans l'espace logique).

SSA+SIA : Toutes choses égales par ailleurs, un observateur devrait raisonner comme s'il était un échantillon aléatoire de l'ensemble de tous les observateurs possibles (séparés de lui dans l'espace-temps ou l'espace logique).[104]

Supposons que je sois *a priori* indifférent devant les possibilités « L'humanité compte $10^{11}$ individus » et « L'humanité compte $10^{14}$ individus ». Je prends maintenant en compte que je suis un être humain, mais je ne prends pas en compte mon rang de naissance. Si mon principe anthropique est SSA+SIA et pas son rival, je dois *a posteriori* croire que la seconde possibilité est extrêmement probable : en piochant au hasard le plus total un individu dans l'ensemble formé par les deux humanités possibles, c'est bien dans la seconde humanité que j'ai le plus de chances de tomber ; cependant, prendre en compte mon rang de naissance rééquilibrerait les probabilités. L'argument de l'Apocalypse n'élève pas SSA+SIA en principe et conclut, après prise en compte du rang de naissance, que c'est la première possibilité qui est nettement plus probable et donc que la fin prochaine de l'humanité est très probable. Toutes ces probabilités se calculent avec l'inférence bayésienne classique. Bostrom est cependant un des premiers à constater que Lewis a sous-estimé la question de la répartition des probabilités sur des possibilités auto-localisantes. Si SIA peut réfuter l'argument de l'Apocalypse, elle conduit à des conclusions fortement contrintuitives dans d'autres problèmes : distribuer des probabilités à la fois dans l'espace logique et dans l'espace-temps ne se fait

[104] Il n'est pas nécessaire de mentionner la classe de référence ici : si la classe est large, elle est certes plus probable (en vertu de SIA), mais cela est compensé par la plus faible probabilité qu'un agent soit *cet agent particulier* au sein de la classe.

pas en suivant un plan rodé. Notons enfin qu'*Anthropic Bias* propose une hypothèse SSSA (*Strong Self-Sampling Assumption*) adaptée aux problèmes où se distinguent des *observer-moments* :

SSSA : Un observateur devrait raisonner comme si son moment présent était un échantillon aléatoire de l'ensemble de tous ses moments dans sa classe de référence.[105]

### 4.3. Stalnaker et le problème de la modélisation de la croyance

Bostrom était un jeune philosophe aujourd'hui célèbre, mais nous pensons qu'il n'a pas remporté un grand succès auprès des épistémologues peu intéressés par le raisonnement anthropique. Celui que tout le monde attendait impatiemment, c'est Robert Stalnaker, un des pères de la Belle au bois dormant, spécialiste de la sémantique des mondes possibles et adversaire de longue date de Lewis. C'est en 2008 qu'il publie *Our Knowledge of the Internal World*, dont un chapitre est consacré à l'auto-localisation. Le professeur du M.I.T. n'est pas qu'un commentateur et un critique des anciens travaux de Lewis, il sait aussi tirer des leçons de certaines jeunes analyses de la Belle au bois dormant, et malgré quelques manques dans la bibliographie du livre, quiconque connait en profondeur la controverse *Sleeping Beauty* comprend que l'auteur s'en inspire partiellement et ne se contente pas de ses premières impressions sur ce problème.

Stalnaker reconnait qu'on peut croire où, quand et qui on est dans le monde, et souhaite esquisser un modèle de croyance auto-localisante, c'est-à-dire un modèle de croyance plus général. Il pense même que les mondes centrés, qu'il envisage comme des parties logiques et temporelles d'un agent, ou des contreparties temporelles, sont indispensables à la

[105] Bostrom (2002), p. 162.

modélisation. Mais il y a un problème : Lewis ne se demande pas si un changement dans la croyance est un changement d'avis ou un changement dans les faits. Cette distinction nous semble justifiée : nous pouvons croire qu'il fait nuit puis croire plus tard qu'il fait jour, soit en admettant une erreur (par exemple nous avions confondu l'obscurité d'un lieu à l'écart de toute source lumineuse avec la nuit), soit parce qu'avec le temps qui passe la nuit a bien laissé place au jour, et que nous le constatons. De même, Lewis ne fait pas la distinction entre une différence de perspective et un désaccord entre agents. Or, dans une conversation, nous avons besoin de comparer et de détacher les croyances des uns et des autres pour nous comprendre, pour savoir où sont accords et désaccords. La critique de Frege par Perry ne change rien au fait que les phrases indexicales, qui renvoient à des contenus différents selon le contexte, gênent la communication. Si le contenu d'une croyance était éternel, était toujours représenté par une classe de mondes non centrés, nous aurions sans doute toujours moyen de nous comprendre ; nous aurions aussi un moyen simple, puisqu'ordinaire, de modéliser la dynamique des croyances d'un agent particulier. C'est alors que Stalnaker, en quelque sorte, déclare : regardez comme nous avons de la chance, car le contenu d'une croyance *est* éternel !

Stalnaker est en fait un eccéitiste. Il connait et emploie le mot quand il paraphrase Lewis, et même s'il ne dit pas explicitement de lui-même qu'il est eccéitiste (le mot ne lui plait peut-être pas), assurément Lewis le qualifierait ainsi[106]. L'état de croyance d'une personne n'est bien décrit que par des mondes centrés, et pourtant le contenu d'une croyance est un ensemble de mondes possibles, non centrés, et ce contenu peut se révéler dans une réflexion, une attention, une conversation. Même si une expérience se fait en deux temps dans le même monde et que le sujet ne sait pas quel temps est *maintenant*. Même dans le cas des deux dieux, bien qu'il

[106] Weatherson (2011) partage cet avis.

tienne de la science-fiction. Dans tout cas « hautement artificiel » où un sujet est incapable de se repérer parmi plusieurs situations possibles qualitativement indiscernables, il reste vrai que le monde où *cette* pensée et *cette* expérience se produisent en lieu et date qui définissent une situation possible est un monde différent du monde où *cette* pensée et *cette* expérience se produisent en lieu et date qui caractérisent une autre situation possible. La réduction eccéitiste au *de dicto* qui marche pour les trois exemples simples précédemment analysés dans ce chapitre (les chemins de campagne, jeudi ou vendredi, et le métro) marche pour ces cas-là aussi. Stalnaker, qui n'est pas certain de convaincre suffisamment, demande au moins au lecteur d'admettre cette réduction pour apprécier le modèle de croyance esquissé.

Un état de croyance est représenté par une paire : un *monde de base*, qui est un monde centré (le croyant, son moment et son monde), et les *mondes de croyance*, un ensemble de mondes centrés (les façons dont le monde pourrait être selon le croyant, et le temps et l'identité qui pourraient être les siens, toujours selon lui). L'information sur les états de croyance d'une série de croyants à divers temps dans divers mondes peut être cryptée par une relation d'accessibilité doxastique, comme dans le modèle standard de Kripke. Formalisé, le modèle de Stalnaker[107] est un sextuplet $< W, S, T, \geq, E, R >$ où :

– W est un ensemble non vide de mondes possibles.

– S est un ensemble de sujets (croyants).

– T est un ensemble d'instants.

– $\geq$ est une relation sur T qui détermine un ordre linéaire des instants (binaire, réflexive, transitive, antisymétrique).

– E est l'ensemble des mondes centrés qui satisfont la condition : le sujet du centre existe dans le monde à l'instant du centre. Sachant qu'un

[107] Stalnaker (2008), p. 69-70.

centre est un couple $< s, t >$ ($s \in$ S, $t \in$ T) et qu'un monde centré est un couple $< c, w >$ ($c$ est un centre, $w \in$ W).

– R est la relation d'accessibilité doxastique sur E (binaire, transitive, euclidienne, sérielle). Dire que $<< s, t >, w >$ R $<< s', t' >, w' >$, c'est dire qu'il est compatible avec ce que le sujet $s$ croit à l'instant $t$ dans le monde $w$, qu'il soit dans le monde $w'$, qu'il soit la personne $s'$ et que l'instant soit $t'$. R satisfait en outre la condition : quels que soient les centres $c$, $c'$ et $c''$ et les mondes $w$ et $w'$, si $< c, w >$ R $< c', w' >$ et $< c, w >$ R $< c'', w' >$, alors $c' = c''$.

La condition imposée sur R est la principale modification faite au modèle lewisien. Intuitivement, elle exige que, pour tout sujet, l'incertitude ou l'ignorance de la place qu'il occupe dans son monde soit toujours incertitude ou ignorance de son monde. Ainsi on peut dire que le contenu d'une croyance est un ensemble de mondes possibles, non de mondes centrés.

Ainsi s'achève une enquête où Lewis est omniprésent, et qui nous a déjà permis de discuter de quelques problèmes d'auto-localisation. Les puzzles vont se faire de plus en plus nombreux à partir de maintenant, jusqu'à ce que le paradoxe ontique-épistémique de la Belle au bois dormant livre des secrets.

# Chapitre 3

# Vers les problèmes compartimentés

Les problèmes que nous appellerons « compartimentés » sont des problèmes d'auto-localisation particuliers. La Belle au bois dormant est l'un d'eux. Avant de discuter longuement de la célèbre énigme probabiliste, nous devons être préparés. Dans un premier temps, nous analyserons Take Five, une charmante énigme qui présente des points communs avec *Sleeping Beauty* mais qui n'est pourtant pas ce qu'il convient d'appeler un problème d'auto-localisation. Pendant cet examen, sera introduit un principe méconnu à la simplicité peut-être trompeuse, qui sera surtout utile plus tard : l'inertie doxastique. Nous nous plongerons ensuite dans l'étude du problème du Prisonnier, difficile puzzle d'auto-localisation dans le temps, mais non compartimenté, apparu très récemment, qui n'a pas de solution unanime. Notre enquête nous permettra de présenter le principe de réflexion, loi importante de la dynamique doxastique. Dans les deux dernières sections, nous définirons le *compartiment* et analyserons neuf problèmes compartimentés, progressant du plus simple vers le plus compliqué. À chaque fois nous essaierons de montrer que le fréquentisme et le bayésianisme s'entendent bien sur les solutions, à tel point que des degrés de croyance pourraient être sans dommage fondés sur des fréquences relatives pertinentes. L'habitude de

ces correspondances ontiques-épistémiques rassurantes ne pourra que rendre le paradoxe de la Belle au bois dormant encore plus inquiétant.

# 1. Take Five et le principe d'inertie doxastique

## 1.1. Le problème Take Five

Bruno adore le jazz et en particulier le célèbre morceau *Take Five*, qu'il écoute régulièrement. Il en possède deux montages différents : une version courte qui dure exactement cinq minutes, et une longue qui dure exactement dix minutes, soit le double. Depuis qu'il a programmé sa chaine hi-fi pour qu'à la demande elle joue au hasard une seule version (supposons équiprobables la lecture de la version courte et celle de la version longue), il ne lance plus que cette lecture aléatoire, à des heures et des jours qui peuvent être très variables. Alice connait les habitudes de Bruno. Elle est capable de reconnaitre le morceau de jazz quand elle l'entend mais ne sait pas distinguer la version courte de la longue lorsqu'elle n'entend qu'un bref extrait, quel qu'il soit. Un jour, Alice décide de rendre visite à son ami. En arrivant, elle entend que *Take Five* sort des haut-parleurs. À quel degré doit-elle croire qu'il s'agit de la version courte ?[108]

Réponse : 1/3. Essayons de le montrer de façon simple. Alice arrive quelque part dans une chaine formée par des intervalles de temps variables où Bruno n'écoute pas son air favori, des intervalles de cinq minutes où il écoute la version courte, des intervalles de dix minutes où il écoute la version longue. Raisonnons ici en fréquentiste, disons, ouvert, pas

[108] Philippe Gay est l'inventeur d'une première version de cette énigme, qu'il n'analyse pas jusqu'au bout. Son blog « Probabilités et énigmes » (http://probas-enigmes.pagesperso-orange.fr/) propose encore d'autres problèmes qui, selon lui, sont analogues à la Belle au bois dormant. Néanmoins, il y a assurément des différences significatives entre la Belle et ces problèmes.

rigoureux à l'excès mais suffisamment prudent. Si la chaine était très longue, de façon à répéter très souvent les écoutes de *Take Five*, il y aurait autant d'écoutes de cinq minutes que de dix, autrement dit deux fois plus d'instants où *Take Five* dix minutes est entendu que d'instants où la version courte est entendue. Alice arrive à un de ces instants, sur lesquels nous supposerons une distribution uniforme des probabilités[109]. Il est normal de considérer qu'elle a deux fois plus de chances d'entendre alors la version longue que d'entendre la courte, et ce quelles que soient ses chances d'entendre un des deux enregistrements et non une autre musique ou le silence : si elle répétait de très nombreuses fois sa visite chez Bruno (nous pouvons les simuler dans une expérience de pensée), une visite sur trois visites où elle est accueillie par *Take Five* serait une visite où elle est accueillie par la version courte. Une probabilité ontique de 1/3 se dégage de cette manière. Pourquoi Alice, qui connait les habitudes de Bruno et est capable de tenir notre raisonnement, lorsqu'elle entend l'air de jazz à son arrivée, n'alignerait-elle pas sur cette probabilité le crédit initial qu'elle associe à « Je suis accueillie par la version courte de *Take Five* » ? Elle a, grâce à ce raisonnement, une raison suffisante pour ne pas être indifférente aux possibilités « version courte » et « version longue », ne pas associer le degré 1/2 à chacune. Tout bien réfléchi, pour « version courte », c'est la probabilité épistémique 1/3 qui s'impose ; 1/2 semble un son discordant, un écart déraisonnable.

Nous pourrions considérer que la visite d'Alice et les habitudes de Bruno forment un grand dispositif aléatoire complexe ayant une certaine propension à réaliser l'écoute de *Take Five* version courte par Alice dès son arrivée, deux fois plus faible que la propension à réaliser l'écoute de la

[109] Nous parlons ici de densité de probabilité. Mais si une telle distribution posait problème, disons au moins que les chances d'Alice d'arriver au moment où se joue la version courte de l'air de jazz sont égales aux chances d'arriver quand se jouent les cinq premières minutes de la version longue, et égales aux chances d'arriver quand se jouent les cinq dernières minutes de cette version.

version longue. Nous le ferions évidemment pour avoir l'occasion d'appliquer le principe principal. Le souci est que Lewis est un incompatibiliste[110] : il tolère qu'on dise d'un instrument aléatoire, tel qu'une pièce de monnaie, qu'il a une propension à faire venir tel résultat, mais pense qu'on ne devrait reconnaitre comme d'authentiques probabilités ontiques que les probabilités irréductibles, c'est-à-dire relative à un phénomène physique primitif, comme la radioactivité. C'est pourquoi un « dispositif aléatoire » Alice-Bruno est pour lui, et pour beaucoup de philosophes, impensable : ce que produit la rencontre de deux volontés humaines est incertain mais ne saurait avoir une « chance ». Nous associons la probabilité ontique (puis épistémique) 1/3 à l'écoute par Alice de la version courte de *Take Five* parce qu'il nous semble nécessaire de nous émanciper *ici* de Lewis. Contentons-nous de notre première explication « fréquentiste », ouverte et simple, si nous voulons répondre à l'interrogation d'Alice.

Maintenant, c'est Bruno qui nous préoccupe. Bruno peut faire quelque chose qu'Alice ne peut pas faire. Au moment où il lance la lecture aléatoire d'une version de sa mélodie préférée, il aligne le crédit initial qu'il accorde à « Ma chaine hi-fi joue la version courte cette fois-ci » sur une chance objective 1/2 assez évidente, que nous pouvons interpréter de façon propensionniste si nous ne sommes pas des incompatibilistes intransigeants. Supposons que version courte et version longue commencent et se terminent par les mêmes phrases musicales et se ressemblent à tel point qu'aucune ne contient de motif qui ne soit pas contenu, une ou plusieurs fois, dans l'autre version ; supposons aussi que Bruno est capable de différencier les versions s'il se concentre sur cette

[110] L'incompatibilisme est d'abord la thèse selon laquelle il ne peut y avoir de libre arbitre dans un univers déterministe. En théorie des probabilités, c'est le refus des probabilités ontiques non extrêmes dans un contexte déterministe. D. Lewis (1980) admet que les phénomènes quantiques sont aléatoires intrinsèquement, qu'ils ne se ramènent pas à un déterminisme. Cf. Cozic et Walliser (2012), p. 14-15 et p. 34-35.

tâche, mais qu'il ne voit plus le temps passer s'il est au contraire charmé par la musique, et que toute différenciation devient alors impossible. Très souvent, Bruno est bercé par *Take Five* pendant toute la durée du morceau. Quand la musique s'arrête, s'il n'a pas noté l'heure exacte de début, il ne peut glaner aucune preuve que telle version vient d'être jouée. Dans ce cas, il doit toujours croire au degré 1/2 que la version courte a été jouée, apparemment en vertu d'un principe de conservation des (degrés de) croyances en absence d'information nouvelle. Mais ce principe est-il aussi connu et simple qu'il en a l'air ? Ouvrons une utile parenthèse.

### 1.2. Le principe d'inertie doxastique

Selon le philosophe Isaac Levi, un principe de conservation concernant les croyances pleines a été défendu en 1877 par le pragmatiste américain Charles Peirce dans *The Fixation of Belief*, un texte qui compare plusieurs méthodes pour se débarrasser de doutes perturbants, atteindre l'état satisfaisant de la croyance (pleine) et surtout *y demeurer* pour en tirer des règles d'action et une volonté d'agir. Tant qu'elle sauve du doute, une croyance n'a pas à être justifiée ; au contraire, un individu *doit* justifier un *changement* de croyance. Levi appelle ce principe « *Doxastic inertia* »[111] et l'énonce de cette manière :

Principe d'inertie doxastique (croyances pleines) : On ne doit pas modifier un état des croyances pleines à moins d'avoir une justification pour le faire.

En 2001, dans son unique article sur la Belle au bois dormant, Lewis suit un principe de conservation concernant les croyances partielles. Mikaël Cozic le relève et l'appelle « inertie doxastique »[112]. C'est une règle

[111] Levi (1994), p. 95.

[112] Cozic emploie l'expression « un principe d'inertie doxastique » lors d'une conférence du cycle PhilEAs à Genève en 2007.

purement épistémologique, c'est-à-dire qui pose sans considération morale des exigences rationnelles quant au réajustement possible de degrés de croyances. D'ailleurs, une telle règle de conservation ne pouvant évidemment pas faire d'une croyance partielle (rendant l'action hésitante) une croyance pleine (sécurisant l'action), ce n'est sûrement pas la doctrine de Peirce qui l'a inspirée. Si l'on désire restituer le vocabulaire de Lewis et de Cozic, cette formulation s'impose :

Principe d'inertie doxastique (croyances partielles) : Tant qu'un agent ne reçoit aucune information *pertinente* relativement à une croyance donnée, il ne doit pas modifier son degré.

De la signification de cette « pertinence » floue dépend notamment la relation entre le principe et la règle de conditionalisation. Lewis ne laisse aucun indice. On repense au chapitre précédent et à ces notions requises pour comprendre le principe principal ; mais relier inertie et résilience, pertinence et admissibilité, c'est se lancer dans des suppositions peu assurées. Un éclairage pourrait venir de Wolfgang Schwarz, qui estime justement que deux écoles, deux points de vue s'affrontent aujourd'hui autour, entre autres, des problèmes d'auto-localisation : l'évidentialisme qui pense qu'une croyance est toujours sur le point d'être modifiée par la moindre preuve, et le conservatisme, selon lequel une croyance est avant tout résistante face à l'afflux d'informations. Schwarz (2015) énonce le principe cher aux « conservateurs » :

Principe du conservatisme doxastique : Si un agent assigne rationnellement un crédit à une proposition A qui ne changera certainement pas de valeur de vérité, alors ce crédit doit rester inchangé tant que l'agent ne reçoit pas une preuve ayant, à la lumière de ses propres croyances, une incidence sur la vérité de A.

Nous reconnaissons un principe d'inertie retouché, mais l'effort entrepris en vue d'évoquer la pertinence relative à une croyance avec des

mots choisis semble insuffisant. Remarquons qu'il n'est question que de croyances éternelles ou assimilées, c'est-à-dire dont on est sûr que le contenu ne changera pas de valeur de vérité au moins pendant une période donnée : dans son article, Schwarz fait référence à la croyance éternelle en un résultat de tirage à pile ou face comme celui de la Belle au bois dormant. Cependant, sa prudente restriction n'est peut-être pas utile, le principe étendu aux croyances temporelles semble parfaitement valable ; en tout cas nous ne voyons pas pourquoi, informé du temps qui passe, nous garderions constant le degré d'une croyance temporelle.

Il semble que la littérature parle peu du principe d'inertie doxastique, dans ces termes ou dans d'autres, et qu'il soit impossible d'y trouver une formalisation. Dans nombre de cas, la « pertinence » n'est pas un problème et la loi se résume sans incident à : pas de nouvelle information, pas de modification des croyances. Ainsi on applique constamment cette règle apparemment simple sans s'en apercevoir, on ne la discute pas, et pourtant tout mouvement doxastique est la conséquence de conditions satisfaites. Quelles conditions exactement ? C'est ce que sont bien obligés de rechercher les analystes des problèmes probabilistes et décisionnels les plus difficiles. La situation de Bruno, lorsqu'il vient d'atteindre la fin de la pièce musicale, ne présente pas une telle difficulté selon nous : personne ne pourrait donner à l'information selon laquelle le morceau vient de s'achever un pouvoir modificateur de la croyance de Bruno selon laquelle il s'agissait de la version courte, sauf si par exemple le morceau a quelque chance de ne pas arriver à son terme quand sa version longue est jouée.

### 1.3. Poursuite de l'analyse de Take Five

Revenons à Alice. Lorsqu'elle rend visite à Bruno et qu'elle entend *Take Five*, elle croit au degré 1/3 que c'est la version courte, alors que Bruno le croit au degré 1/2. Cela est normal : les deux agents ne disposent

pas tout à fait des mêmes données. On ne peut pas dire qu'Alice se localise dans le temps, bien que dans notre raisonnement elle se localise dans la chaine des occupations quotidiennes de son ami. Alice doit considérer qu'à son arrivée, si elle entend *Take Five*, c'est soit les cinq minutes de la version courte, soit les cinq premières minutes de la longue, soit les cinq dernières minutes de la longue. Cette égalité des trois durées garantit l'équiprobabilité des trois possibilités. Son espace des observables (univers) contient trois objets, alors que celui de Bruno, qui ne voit pas le temps passer, contient deux objets seulement : il confond les cinq premières et les cinq dernières minutes de la version longue. Ce qui est intéressant, c'est d'examiner ce qui se passe quand nos deux amis se rencontrent. Nous pressentons qu'ils ne peuvent pas rester en désaccord sur leurs degrés de croyance.

Alice aura beau parler avec Bruno dès son arrivée, rien ne fera en sorte qu'elle modifie le crédit 1/3 si elle ne reçoit aucune information dont il dépend. Il en est de même pour Bruno ; c'est pourtant lui qui doit changer ses croyances, et il ne doit pas le faire parce qu'il sait qu'Alice croit différemment, il doit le faire parce que l'arrivée d'Alice au moment où il écoute *Take Five* est une information pertinente. En effet, il est clair que plus longtemps dure la lecture de la musique et plus probable est l'arrivée d'Alice entre-temps. Supposer une distribution uniforme des probabilités dans le temps nous conduit à penser que la probabilité conditionnelle de la lecture de la version courte sachant qu'Alice arrive entre-temps est deux fois moindre que la probabilité conditionnelle de la lecture de la version longue sachant la même donnée. Le glissement bayésien fait le reste : Bruno doit diminuer jusqu'à 1/3 le crédit qu'il accorde à la lecture de la version courte dès qu'Alice met un terme à son évasion musicale. Ainsi les degrés rationnels de croyance des deux sujets s'harmonisent avant même qu'ils en discutent ensemble. Remarquons alors cette curiosité : Bruno pouvait être perturbé par n'importe quoi, par un

chien qui aboie à proximité, ou encore par la sonnerie du téléphone. Il plonge en apnée musicale en croyant équiprobables les possibles « version courte » et « version longue », il en ressort en brisant cet équilibre. Il faut comprendre la raison : il a plus de chances d'arriver au bout de la version courte (sans être dérangé) que d'arriver au bout de la longue. Il prend en compte le temps, des quantités de temps, pour modifier ses croyances. Il pourrait se localiser dans le temps et probabiliser des parties temporelles de lui-même, et pourtant, parce que sa méconnaissance du monde est manifeste (il ignorait s'il serait dérangé, et par quoi), il semble plus naturel qu'il n'envisage que des mondes possibles, des mondes où il est dérangé dans son écoute de la version courte, ou bien dans son écoute des cinq premières minutes de la longue, etc.

La différence avec la Belle au bois dormant est là : Bruno sera ou ne sera pas interrompu pendant son écoute, et s'il l'est c'est une seule fois, qu'il ne peut pas oublier ; une fois revenu de son évasion musicale, il est certain qu'un retour similaire n'a pas eu lieu quelques minutes auparavant, donc il n'est pas obligé de se localiser dans le temps, d'avoir des croyances autres que *de dicto*. L'information qu'il reçoit quand son écoute est interrompue est elle-même *de dicto*, éternelle. Des raisonnements classiques, éprouvés et sûrs le conduisent à modifier aisément ses croyances. Pour qu'un paradoxe apparaisse, il faudrait imaginer un scénario un peu différent, avec plus de malice. Nous en proposons un ici, non pour résoudre le problème engendré (nous ne sommes pas encore prêts), mais pour montrer comment Take Five peut entrer dans la famille des énigmes analogues à la Belle au bois dormant.

*Take Five* dix minutes est très semblable à la version de cinq minutes. Alice et Bruno savent que cette version longue est en outre extraordinaire : des notes à effet hypnotique sont jouées durant la cinquième minute, et durant la sixième minute la personne qui les a entendues oublie tout ce qui lui est arrivé pendant les cinq dernières minutes écoulées, si bien qu'elle a

l'impression d'écouter la première minute du morceau. Pour la personne hypnotisée, la musique est continue et semble ne durer que cinq minutes, rien ne permet de distinguer un avant et un après hypnose. Alice et Bruno lancent une lecture aléatoire : *Take Five* cinq minutes sera lu avec probabilité 1/2, *Take Five* dix minutes avec probabilité 1/2. Bruno n'a pas de montre et est seul à écouter l'air de jazz dans son casque ; sans casque, Alice n'entend rien, peut consulter sa montre, observer son ami, l'interrompre pour lui poser une question avant qu'il ne se replonge dans la musique. Bruno sait qu'Alice va l'interrompre dans tous les cas lors de la troisième minute de l'écoute, et en cas de version longue également lors de la huitième minute, pour lui demander la probabilité qu'il écoute la version courte. Quelle serait la réponse la plus rationnelle ?

Cette variante efface la différence avec la Belle au bois dormant que nous avions précédemment notée. Ici, dans le cas où la version longue est jouée, il n'y a pas deux fois plus de chances qu'il y ait une unique interruption, mais il y a deux fois plus d'interruptions, *sans que Bruno puisse les compter*. Celui-ci connait le protocole de l'expérience et *apparemment* n'apprend aucune information neuve lors d'une interruption. Aussi il est difficile de reconstituer un espace des observables et de pencher vers la réponse 1/2 ou la réponse 1/3. L'analogie avec la Belle est peu discutable. Sans doute, ce Take Five avec hypnose est une énigme d'auto-localisation qui ne peut pas être résolue sans entrer dans des développements longs et des débats profonds, dont nous verrons la teneur quand nous regarderons la Belle au bois dormant bien en face.

## 2. Le Prisonnier et le principe de réflexion

Dans un article de 2003, « *Some Problems for Conditionalization and Reflection* », Frank Arntzenius énonce quelques problèmes d'auto-

localisation déconcertants, et notamment l'énigme du Prisonnier[113]. Pour apprécier celle-ci pleinement, nous pensons avec Arntzenius qu'il faut préalablement présenter le principe de réflexion, une des grandes règles toujours controversées de l'épistémologie bayésienne.

### 2.1. Le principe de réflexion de van Fraassen

Bas van Fraassen propose ce principe en 1984 dans l'article « *Belief and the Will* », au sein d'une thèse volontariste que Pascal Engel (1995) résume en quatre points :

> (i) toute croyance que *p* est l'expression d'un engagement subjectif envers la vérité de *p* (toute croyance implique une forme d'acceptation),
>
> (ii) cet engagement implique que nous prenions en compte non seulement notre croyance en la vérité présente ou passée de *p*, mais aussi notre croyance future en la vérité de *p*,
>
> (iii) et le degré de notre croyance présente en la vérité de *p* doit refléter le degré de notre croyance future en la vérité de *p* (Réflexion),
>
> (iv) par conséquent nos croyances futures contraignent nos croyances présentes.

Selon van Fraassen, même si un agent peut juger par exemple que « Il ne pleuvra pas demain soir » ne contredit pas formellement « Je croirai demain matin au degré 0,3 qu'il pleuvra le soir », il n'est pas rationnel qu'aujourd'hui il croie vraies en même temps ces deux propositions. Si la seconde est pour lui certaine, alors il doit croire *maintenant* au degré 0,3 qu'il pleuvra demain soir. Il ne doit s'autoriser aucun autre degré de croyance et notamment ne doit pas être certain qu'il ne pleuvra pas demain soir, mais cette incertitude est paradoxalement la marque d'une « confiance

[113] À ne pas confondre avec le dilemme du prisonnier, un des grands problèmes de la théorie des jeux, ni avec le paradoxe des (trois) prisonniers, un problème proche du plus célèbre Monty Hall.

en soi » et d'un « optimisme »[114] authentiques puisqu'elle est fondée sur la certitude d'une incertitude future (qui plus est quantifiée). Bien entendu, si un individu est certain qu'il sera, dans l'avenir, certain que A, c'est maintenant qu'il doit croire pleinement que A. Un agent rationnel et confiant respecte donc cette loi de bon sens :

Principe de réflexion : $C_t(A \mid \langle C_{t+x}(A) = r\rangle) = r$

$C_{t+x}(A)$ est le crédit que l'agent accorde à une proposition A à un moment postérieur au moment $t$ de la réflexion. Au temps $t$, lorsque l'agent prédit avec certitude $\langle C_{t+x}(A) = r\rangle$, il doit accorder le crédit $r$ à A. Il va sans dire que van Fraassen n'envisage que des propositions éternelles. Remarquons la ressemblance avec le principe de Miller et le principe principal : la réflexion est un principe d'alignement, mais celui-ci est diachronique alors que l'alignement entre probabilités ontique et épistémique est synchronique (on peut, abusivement ou pas, l'affirmer aussi pour le principe principal si l'on remarque que seule la chance ou propension est datée).

En 1995, dans « *Belief and the Problem of Ulysses and the Sirens* », van Fraassen corrige et généralise son principe. Après analyse des critiques de plusieurs auteurs[115], il convient que la réflexion ne s'applique que si les croyances futures sont rationnelles : si un agent peut prédire sa folie passagère tel Ulysse assuré de bientôt succomber au chant des sirènes, il serait irrationnel qu'il applique la réflexion. Si je sais que j'ai pris une drogue qui me convaincra dans une heure que je peux voler comme Superman, je ne dois pas croire présentement que je peux voler. En ce qui concerne la généralisation, van Fraassen convient qu'un ensemble de valeurs prédictibles peut et doit être « reflété » dans le présent :

[114] Van Fraassen (1984), p. 244. « Confiance en soi » et « optimisme » sont d'autres noms que l'on peut donner au principe de réflexion, selon l'auteur.

[115] Parmi eux, Richard Jeffrey, William Talbott, David Christensen et Patrick Maher.

Principe de réflexion généralisé : $C_t(A)$ doit appartenir au plus petit intervalle couvrant toutes les valeurs prédictibles de $C_{t+x}(A)$.

La relation entre conditionalisation et réflexion n'est pas claire. Van Fraassen estime d'abord que le principe de conditionalisation implique celui de réflexion, et que le second peut être considéré comme une alternative au premier. Plus tard, il défend l'implication dans le sens inverse, en précisant que si celle-ci est avérée la réflexion offre à la conditionalisation une nouvelle justification[116]. Cependant, dans un papier récent, Jonathan Weisberg réfute les arguments en faveur de ces implications[117]. Des analystes des problèmes d'auto-localisation (parmi lesquels Arntzenius, analyste du Prisonnier et de la Belle au bois dormant entre autres) notent que certains scénarios appellent une transgression à la fois de la réflexion et de la conditionalisation, mais cela ne les empêche pas d'évoquer aussi la réflexion comme si elle avait un statut autonome par rapport à la conditionalisation.

Le principe de van Fraassen n'est peut-être pas comparé et connecté au bon candidat. Certes, il produit un mouvement doxastique, une révision de $C_t(A)$ lorsque l'agent se considère comme informé par $\langle C_{t+x}(A) = r \rangle$. Mais en préliminaire de cette révision il y a un regard tourné vers le futur, une anticipation de mouvements doxastiques (ou d'absence de mouvements) qui conduiront au crédit futur de A ; en conclusion de cette anticipation, il y a une prédiction de la valeur (ou des valeurs possibles) de $C_{t+x}(A)$ et sur elle un réajustement de $C_t(A)$ par lequel l'agent espère garantir que ce crédit sera *conservé* entre $t$ et $t + x$. Regardée sous cet angle, la réflexion est une inertie doxastique à l'envers : le principe d'inertie doxastique réactualise un degré de croyance rappelé du passé, après constatation d'une absence d'informations susceptibles de le modifier ; le

[116] Van Fraassen (1999).

[117] Weisberg (2007).

principe de réflexion hâte un degré de croyance futur prédit et fait en sorte qu'aucune information prévue entre-temps ne sera capable de le modifier. Réflexion d'un crédit futur et conservation d'un crédit passé apparaissent comme des actes rationnels complémentaires et indissociables ; si notre mémoire du passé et notre anticipation de l'avenir étaient parfaites, nous les confondrions peut-être en un seul et même acte, ou ne les distinguerions plus qu'en discernant aussi passé et avenir, mémoire et anticipation. Nous pouvons à présent davantage regretter le manque d'études du principe d'inertie doxastique qui nous auraient permis d'appuyer cette thèse téméraire.

### 2.2. Le problème du Prisonnier

Maintenant que nous connaissons mieux le principe de réflexion, parlons du problème d'Arntzenius. Le prisonnier, agent rationnel, est enfermé à 18 heures. Le gardien, un homme de parole, annonce qu'il va secrètement tirer à pile ou face : si et seulement si face, il éteindra l'unique lampe de la cellule à minuit. Comme il n'y a pas de pendule dans la cellule, le prisonnier a du mal à apprécier l'écoulement du temps. Quelques heures passent. La lumière est toujours là. Il est peut-être plus de minuit, rien n'est sûr. Selon Arntzenius, à ce stade le degré de la croyance du prisonnier en l'obtention de pile doit augmenter ; gageons que beaucoup de philosophes sont d'accord. En effet, à la place du prisonnier nous aurions tendance à probabiliser grossièrement notre position temporelle. Plus le temps passe, plus il devient probable puis certain qu'il est plus de minuit. Supposons que la cellule reste éclairée : il est contraire à l'intuition que le crédit 1/2 accordé initialement à pile devienne brutalement 1 seulement lorsqu'il est certain que minuit est passé. Le passage de 1/2 à 1 est intuitivement progressif ou continu. Dépassons l'intuition, osons mêler les localisations dans l'espace logique et dans le temps. Si nous ne sommes pas certains que

minuit est passé, nous sommes face à trois possibilités centrées, exclusives et conjointement exhaustives :

FM : « La pièce tombe sur **f**ace et il est **m**oins de minuit »

PM : « La pièce tombe sur **p**ile et il est **m**oins de minuit »

PP : « La pièce tombe sur **p**ile et il est **p**lus de minuit »

Associer rationnellement une probabilité à ces possibilités, qui plus est une probabilité amenée à changer au cours du temps, est une tâche extrêmement difficile. Pourtant, nous savons que le crédit de PP va passer de 0 à 1 si la lampe reste allumée, tandis que celui des deux autres possibilités va passer de 1/2 à 0. Nous ne pouvons pas comprendre ces changements sans leur accorder un caractère progressif, et nous croyons que le crédit de PP va augmenter plus vite que ne diminuera le crédit de PM, puisque celui-ci n'est pas seul à tendre continument vers 0 (il y a aussi le crédit de FM) et que la somme des trois crédits doit être 1 à tout moment. Si l'hypothèse éternelle « La pièce tombe sur pile » est bien PM $\vee$ PP, sa probabilité doit donc progressivement augmenter jusqu'à 1. Ainsi les croyances temporelles semblent affecter les croyances éternelles, mais en essayant d'expliquer comment, nous balbutions. L'intensification de la croyance en pile ne peut être ruinée que par une extinction des feux, qui fournit clairement l'information « la pièce tombe sur face et il est (plus de) minuit », donc annule immédiatement le crédit accordé à pile. Mais quelle information ou quelles informations ont confirmé pile quand la lampe brillait ?

Épaississons le mystère. Dès 18 heures, le prisonnier est capable de prédire l'intensification future de sa croyance en pile : il sait que lorsque minuit approchera, le degré de cette croyance ne sera certainement plus 1/2, parce que sa cellule sera encore éclairée et qu'il ne sera pas certain que minuit n'est pas encore passé. D'après Arntzenius, si à 18 heures le

prisonnier croit en pile au degré 1/2, il viole le principe de réflexion. Comme il semble que 1/2 est le degré le plus rationnel à cette heure-là, qu'un degré futur supérieur est lui aussi rationnel et qu'en plus aucune information éternelle prédictible (à 18 heures) ne sera reçue entre-temps, nous pourrions en effet penser que ce scénario fournit un contre-exemple au principe de van Fraassen.

### 2.3. Une solution possible

Darren Bradley, prolifique analyste de la Belle au bois dormant, estime que l'énigme du Prisonnier est moins hermétique qu'il n'y parait. Sa solution est la suivante[118] : le prisonnier reçoit contre toute attente des informations éternelles qui confirment pile. En voici un exemple. Supposons qu'à 18 heures il croit au degré 0 qu'il est plus de minuit et qu'à 23 heures il le croit à un degré supérieur à 0,3. À 23 heures, il ne sait pas qu'il est 23 heures, mais il voit la lampe allumée et sait par une bonne introspection et ce qui lui reste de notion du temps que le crédit qu'il accorde à « il est plus de minuit » est supérieur à 0,3. Ces résultats se combinent dans une nouvelle croyance éternelle $A_{0,3}$ : « La lampe est allumée après que le crédit que j'accorde à « il est plus de minuit » a augmenté jusqu'à dépasser 0,3 ». Le prisonnier ignorait $A_{0,3}$ à 18 heures, il ne pouvait pas prédire avec certitude que la lampe serait allumée. Et cela est vrai pour d'innombrables crédits autres que 0,3. Les $A_x$ sont des informations sur lesquelles le prisonnier peut conditionaliser. La probabilité subjective de $A_{0,3}$ sachant pile est évidemment 1, elle est strictement supérieure à la probabilité de $A_{0,3}$ sachant face (en cas de face, la lampe ne reste pas allumée toute la nuit). Cela signifie que $A_{0,3}$ confirme pile. Ainsi

[118] Bradley (2011b), sections 3 et 4. Bradley (2007), p. 56-62, expose une première version de cette solution. Préférons la version la plus récente.

ce ne sont pas seulement des croyances temporelles qui ont affecté la croyance en pile, mais surtout la découverte d'une preuve.

Il apparait donc que le prisonnier ne viole pas le principe de réflexion à 18 heures. Lorsqu'il anticipe un moment où il intensifie sa croyance en pile, il envisage en réalité un futur possible, l'autre possibilité étant qu'il croit en pile au degré 0. En effet il ne peut pas prédire avec une certitude absolue l'état de la lampe. Si nous considérons qu'il y a plusieurs valeurs prédictibles pour le crédit de pile à un instant futur donné, le plus petit intervalle couvrant ces valeurs est nécessairement $[0 ; x]$ où $1/2 < x \leq 1$. Or, le crédit 1/2 à 18 heures appartient à cet intervalle, ce qui signifie que le prisonnier respecte le principe de réflexion généralisé.

Remarquons que la réponse de Darren Bradley a un curieux accent eccéitiste, bien que ce philosophe inspiré par Lewis et Perry ne partage ni les thèses de Stalnaker ni la solution de la Belle au bois dormant proposée par celui-ci. Bradley ne ramène pas des croyances temporelles à des croyances éternelles équivalentes ; en revanche, il explique que la « mutation des croyances temporelles » n'est qu'indirectement une cause de l'intensification de la croyance (éternelle) en pile, et que le prisonnier conditionalise sur une croyance éternelle découverte. Il distingue la *découverte*, « modification de croyance en vertu de la découverte de la vérité du contenu de la croyance, dont la valeur de vérité n'a pas changé durant la période considérée », et la *mutation*, « modification de croyance en vertu d'un changement de la valeur de vérité du contenu de la croyance »[119]. Une information temporelle est souvent acquise par mutation : par exemple, j'apprends qu'il est 9 heures en regardant ma montre, alors qu'il n'était pas 9 heures quand je l'ai consultée peu de temps auparavant. Son examen des problèmes d'auto-localisation incite Bradley à penser qu'une information temporelle acquise par mutation n'est jamais

[119] *Ibid.*, section 2.

pertinente relativement à une hypothèse éternelle, c'est-à-dire ne peut pas modifier le degré de la croyance d'un quelconque agent en cette hypothèse. Le prisonnier *semble* trouver de la pertinence dans une telle information, mais en réalité c'est bien une information éternelle découverte, telle que $A_{0,3}$, qui confirme pile. Un eccéitiste s'efforce de trouver une croyance propositionnelle cachée derrière une croyance auto-localisante, notamment parce qu'évaluer le degré d'une croyance propositionnelle n'est pas étranger à ses habitudes. De façon assez similaire, on peut s'efforcer de découvrir, comme Bradley, une information éternelle cachée derrière une information temporelle apparemment pertinente relativement à une hypothèse éternelle, notamment pour réviser la croyance de manière classique.

Ainsi, que ce soit dans le problème du métro pour l'IHPST[120] ou ici avec le Prisonnier, les estimations de probabilités épistémiques les moins simples (probabilité d'arriver à Mabillon plutôt qu'à Odéon, probabilité de pile légèrement confirmée par une lampe allumée…) tentent d'écarter les lieux, les instants, les identités possibles, pour ne garder que les mondes possibles. Tout se passe comme si on se raccrochait aux mondes et à un dénombrement salvateur, comme si on cherchait une proportion de mondes ayant une certaine propriété. Avec l'écoulement du temps, les auto-localisations possibles deviennent impossibles et inversement, un compte des possibles est remplacé par un autre, un espace des observables par un autre. Les estimations difficiles recherchent plus de constance, de stabilité. C'est aussi l'aveu que le bayésianisme se sent parfois inefficace sans un support objectif, si ce n'est ontique. La tentation de se demander « Quelles sont mes *chances* de me trouver dans la bonne situation, celle où j'arrive à Mabillon, celle où ma cellule reste éclairée ? » n'a jamais été aussi grande

[120] Le scénario du métro est brièvement analysé dans le chapitre 2 de cette thèse.

que dans l'analyse de ces problèmes délicats qui, pourtant, ne semblent pas *a priori* s'adresser à des objectivistes, fréquentistes ou propensionnistes.

## 3. Les problèmes compartimentés

### 3.1. Définition

Le prisonnier ne fait pas que découper sa nuit en « avant minuit » et « après minuit » : pour penser une évolution continue des probabilités qu'il estime, il envisage nécessairement de très nombreux moments, voire, s'il est lewisien, de très nombreux centres, parties temporelles de lui-même, et en outre il les délimite mal. Bien que mesurant le temps avec difficulté, il situe encore un avant et un après, passe continument d'une position à une autre et ne revient jamais sur un « temps révolu », parce qu'il ne souffre pas d'amnésie. Ces particularités ne se retrouvent pas dans ce que nous appelons les *problèmes compartimentés*, dont fait partie notre dernière variante de Take Five. Les problèmes compartimentés sont bien des problèmes d'auto-localisation dans *le* monde (dans le temps précisément) et/ou dans *les* mondes (dans l'espace logique traditionnel), mais leur difficulté ne vient pas du fait que l'agent serait gêné par des estimations probabilistes évoluant continument, mais du fait qu'il est gêné par une mémoire défaillante ou un incident cognitif dû, par exemple, à une drogue ou, comme dans le cas du mélomane Bruno, à l'effet hypnotique d'une musique.

L'agent vit une expérience où il *sait* qu'il devient partiellement amnésique à certains moments, si bien que, incapable de se situer avec exactitude dans une période (de temps) ou dans une autre, il a besoin d'estimer comme possiblement actuels un nombre précis de centres souvent colocalisés (dans un même monde). Appelons *compartiment* un monde

centré (ou un ensemble de mondes centrés) possiblement actuel du point de vue du sujet amnésique et que celui-ci, à moins d'être (ré)informé, est incapable de reconnaitre, de distinguer des autres mondes centrés pourtant disjoints que le protocole de l'expérience lui fait envisager comme possibles. Par exemple, quand Bruno est interrompu par Alice lors de ce qui lui *semble* être la troisième minute de l'écoute de son air favori, il ignore quelle version de *Take Five* est jouée et ignore s'il a été hypnotisé quelques minutes avant, c'est-à-dire s'il est dans la huitième minute de l'écoute de la version longue plutôt que dans la troisième minute de l'écoute d'une des deux versions ; il a donc besoin de probabiliser trois compartiments :

– < < Bruno, 3e minute de l'écoute >, version courte >

– < < Bruno, 3e minute de l'écoute >, version longue >

– < < Bruno, 8e minute de l'écoute >, version longue >

Bruno n'a pas besoin d'un autre découpage avec des possibles plus nombreux. Comme le premier compartiment possède un paramètre (le centre, ou le temps plus précisément) commun avec le deuxième compartiment mais pas avec le troisième, et comme le deuxième compartiment possède un paramètre (le monde) commun avec le troisième compartiment mais pas avec le premier, on pressent que le calcul des probabilités associées sera simple dans le sens où il mènera à des valeurs comme 1/2, 1/3 ou 1/4, et que ce sont les réflexions philosophiques en amont du calcul qui vont l'orienter et déterminer ces probabilités.

### 3.2. Généralisation des expériences de type Belle au bois dormant

Mettons de côté Take Five et préparons-nous à étudier plusieurs problèmes compartimentés aux scénarios très similaires puisqu'ils sont tous des variations de la Belle au bois dormant. Nous allons pour cela tenir

compte de plusieurs variantes proposées dans la littérature et notamment le *generalized Sleeping Beauty problem* inventé par Roger White[121]. Dans ce qui suit, un « réveil » désigne un réveil similaire à ceux de l'expérience originale : les scientifiques sortent la Belle de son sommeil profond vers midi, ont un bref entretien avec elle, puis la rendorment avec un somnifère spécial qui lui fait oublier tout ce qui s'est passé durant ce réveil.

La Belle participe à des expériences dont nous allons schématiser les protocoles afin de les comprendre au premier coup d'œil. Les scientifiques disposent de deux générateurs aléatoires très fiables : le premier, appelé G1, renvoie au hasard le plus total un nombre entier compris entre 1 et $n$ (inclus) lorsqu'on entre le paramètre entier positif non nul $n$ dans son programme ; le second, G2, renvoie 1 avec probabilité $x$ ou 0 avec probabilité 1-$x$ lorsqu'on entre le paramètre $x$, réel de l'intervalle unité, dans son programme. Les scientifiques endorment la Belle, puis procèdent au Tirage du dimanche soir : ils utilisent G1, réglé sur la limite haute $n$, pour obtenir aléatoirement un entier. Durant $p$ jours consécutifs, ils font subir à la Belle une série, fonction du résultat du Tirage, de réveils possibles, l'effectivité de chacun d'eux étant décidée par une consultation de G2 ajusté sur la probabilité de réveil désirée. Le $(p+1)^{ème}$ jour, la Belle se réveille et l'expérience est terminée. Précisons que la Belle connait dès dimanche $n$, $p$, toutes les probabilités des réveils et l'ensemble du protocole. Ce dernier peut être schématisé par un tableau de $n$ lignes et $p$ colonnes donnant les probabilités des réveils. Cependant, il est possible d'introduire une règle modifiant un peu l'organisation du tableau : dans une série, une possibilité de réveil peut prendre place, non pas au milieu de la période élémentaire d'un jour, mais au milieu d'un jour dans une période de deux ou plusieurs jours consécutifs ; dans ce cas, si le réveil doit avoir lieu (c'est-à-dire si G2 renvoie 1), équitablement on tire au sort le jour du

[121] White (2006).

réveil (en réutilisant éventuellement G1), et aucun autre réveil ne peut avoir lieu durant les autres jours de la période. Schématisons et expliquons un exemple d'expérience avec $n = 3$ et $p = 4$ :

| Jour / Tirage | **Lundi** | **Mardi** | **Mercredi** | **Jeudi** |
|---|---|---|---|---|
| **1** | 1/3 | 0 | 1 | 0 |
| **2** | 3/4 | | 1/4 | 1/4 |
| **3** | 0 | | | |

Ici, le résultat du Tirage (du dimanche soir) est 1, 2 ou 3. Si c'est 1, la Belle est réveillée lundi avec probabilité 1/3 (reste endormie avec probabilité 2/3) ; elle est réveillée assurément mercredi ; mardi et jeudi elle dort toute la journée. Si le résultat du Tirage est 2, la Belle est réveillée avec probabilité 3/4 dans la période lundi-mardi, soit le lundi, soit le mardi ; elle est réveillée avec probabilité 1/4 mercredi et aussi avec probabilité 1/4 jeudi. Si le résultat est 3, la Belle dort pendant toute la durée de l'expérience. Notons que si celle-ci était répétée une infinité de fois, le nombre de réveils par expérience serait en moyenne 31/36, qui est la somme des probabilités du tableau divisée par $n$. Lorsque la Belle se réveille durant une expérience, elle est certaine que G1 n'a pas généré 3 dimanche soir mais ne peut pas se prononcer sur les issues 1 et 2 ; elle ignore totalement quel jour on est car, bien que le protocole interdise le réveil les mardi et jeudi dans le cas de l'issue 1, il n'en est pas de même dans le cas où 2 aurait été généré. Finalement, les cases du tableau marquées d'une probabilité non nulle correspondent aux compartiments envisagés par l'agent. Un cas particulier est la double case marquée avec la probabilité 3/4 : même si la Belle savait que 2 a été généré dimanche et que le jour présent n'est ni mercredi ni jeudi, elle ne pourrait toujours pas

savoir si on est lundi ou bien mardi, aussi elle doit séparer les compartiments < < Belle, lundi >, 2 > et < < Belle, mardi >, 2 >.

## 4. Neuf problèmes de type Belle au bois dormant

Maintenant que nous sommes capables de schématiser la plupart des expériences de type Belle au bois dormant, nous pouvons nous préparer à affronter la redoutable énigme originale grâce à l'analyse préalable de neuf problèmes beaucoup plus simples. Nous nous efforcerons à chaque fois d'envisager les points de vue bayésien et fréquentiste afin de constater l'identité de leurs résultats.

### 4.1. Problème n° 1

L'expérience de ce premier problème peut être schématisée ainsi :

| Jour / Tirage | **Lundi** |
|---|---|
| **1** | 1 |
| **2** | 1 |

L'unique question posée par les scientifiques à la Belle qui se réveille en cours d'expérience est la suivante : « Quelle est la probabilité que le Tirage ait amené 1 ? »

S'il suffit, pour que l'on ait des compartiments, que la Belle hésite entre deux classes de mondes possibles et perde la mémoire de son réveil du lundi, alors ce problème est compartimenté puisque l'agent envisage en cours d'expérience les compartiments < < Belle, lundi >, 1 > et < < Belle, lundi >, 2 > qui correspondent aux deux cases du tableau

marquées d'une probabilité. Si nous considérons que la notion de compartiment est indissociable de celle de perte d'information temporelle, alors ce problème n'est pas compartimenté puisque le seul moment où l'agent a une petite difficulté à se repérer dans le temps est le mardi, jour où il doit se rendre compte que l'expérience est terminée et que la perte de mémoire lui a donné l'impression brève d'être réveillé lundi. Dans tous les cas, le problème est classique dans le sens où la Belle n'a pas besoin de se localiser dans le temps pour évaluer les probabilités des résultats possibles du Tirage ; l'amnésie et la distinction entre une expérience en cours et une période hors-expérience sont réduites à des détails scénaristiques de peu d'intérêt.

La réponse de la Belle à la question des scientifiques est à l'évidence 1/2. Choisir une interprétation de la probabilité plutôt qu'une autre ne change rien. Le bayésien argumenterait ainsi : la propension de G1 à générer 1 est 1/2 dès que le paramètre $n = 2$ est entré dans son programme ; donc, en vertu du principe principal, la Belle doit croire dimanche au degré 1/2 que le Tirage va amener 1. Lorsqu'elle se réveille lundi, événement qui devait assurément arriver, elle n'apprend rien de pertinent relativement à sa croyance. Elle « apprend » pour ainsi dire qu'on est lundi, mais il ne s'agit que du constat du temps qui passe et de la mutation normale, indépendante du Tirage, de la valeur de vérité associée à « On est lundi ». C'est pourquoi, en vertu du principe d'inertie doxastique, elle doit toujours croire au degré 1/2 que le Tirage a amené 1. Le fréquentiste, qui a notamment dans son arsenal la loi des grands nombres, tiendrait quant à lui ce raisonnement : dans une série longue voire infinie d'expériences répétées, la fréquence des 1 parmi tous les entiers générés par G1 serait 1/2, étant donné que l'espace des observables le plus approprié, {1, 2}, ne contient que deux éléments dont aucun détail de l'expérience ne remet en cause l'équiprobabilité. La Belle serait elle-même questionnée (sur la probabilité de l'issue 1) une fois sur deux lorsque 1 a effectivement été généré, une

fois sur deux lorsque 2 a été généré. L'événement singulier « G1 génère 1 lors de *cette* expérience » a donc pour probabilité 1/2.

### 4.2. Problème n° 2

Voici le tableau qui résume la nouvelle expérience :

| Jour / Tirage | **Lundi** |
|---|---|
| **1** | 1/2 |
| **2** | 1/2 |

L'unique question posée par les scientifiques à la Belle qui se réveille en cours d'expérience est encore une fois : « Quelle est la probabilité que le Tirage ait amené 1 ? »

Ce problème ressemble assez au précédent. La différence réside dans l'incertitude d'un réveil le lundi. La Belle est dans cette incertitude le dimanche avant d'entrer dans l'expérience, ainsi que le mardi lorsqu'elle a possiblement été réveillée lundi mais ne peut pas s'en souvenir. Si elle se réveille lundi, elle apprend justement qu'elle est réveillée lundi, et il s'agit d'une information éternelle qui aurait pu être pertinente pour l'issue du Tirage… Mais ici la probabilité (évaluée dimanche) d'un réveil le lundi est 1/2 : c'est la même pour les deux issues possibles du Tirage. Il faut donc considérer que la décision de G2 de sortir ou pas la Belle de son sommeil est indépendante du résultat du Tirage qui utilise G1. Un bayésien dirait donc que l'information nouvelle du lundi manque de pertinence et que ce jour-là la Belle doit croire au degré 1/2 que l'issue du Tirage est 1, degré calqué sur celui du dimanche par respect du principe d'inertie doxastique. De toute façon, conditionaliser sur « La Belle est réveillée lundi » ne conduit à aucune variation de la croyance en l'issue 1. Remarquons que la

Belle engagée dans une expérience singulière aura peut-être l'occasion d'être informée d'un réveil lundi, mais n'aura certainement pas l'occasion d'être informée de son éventuel non-réveil, de son sommeil ininterrompu lundi. Pour être informé il faut être conscient ; cela signifie ici que l'état conscient de l'agent lundi est à la fois condition du savoir et objet du savoir. Dans le problème original de la Belle au bois dormant, cela a son importance.

Un fréquentiste tiendrait un raisonnement similaire à celui du problème n° 1. La seule difficulté réside dans le fait qu'une expérience ne produit pas toujours un réveil lundi, qu'il soit un réveil-1 (c'est-à-dire dans un monde où 1 est le résultat du Tirage) ou un réveil-2. Pourtant, que le fréquentiste détermine la fréquence des 1 parmi tous les entiers générés par G1 ou bien la fréquence des réveils-1 parmi tous les réveils, le résultat sera le même : 1/2. Et 1/2, c'est aussi la probabilité bayésienne. Là encore, la probabilité épistémique du bayésien et la probabilité ontique du fréquentiste ont des valeurs identiques. Cette identité n'est pas l'effet d'un choix particulier de la probabilité du réveil lundi : cette probabilité aurait pu être 1/4 ou 5/6, sans rien changer à la réponse 1/2 d'une Belle fréquentiste ni à la réponse 1/2 d'une Belle bayésienne. Plus généralement, lorsque les lignes du tableau présentent des séries de possibilités de réveils identiques, on ne voit pas pour quelle raison la Belle, qu'elle soit fréquentiste ou bayésienne, préférerait une issue du Tirage à une autre.

### 4.3. Problème n° 3

Présentons un autre protocole d'expérience :

| Jour / Tirage | **Lundi** |
|---|---|
| **1** | 1/2 |
| **2** | 1 |

La question posée par les scientifiques à la Belle qui se réveille en cours d'expérience est toujours : « Quelle est la probabilité que le Tirage ait amené 1 ? »

Comme dans les deux précédents problèmes, lundi est le seul jour de l'expérience à proprement parler, c'est-à-dire le seul jour au milieu duquel peut avoir lieu le réveil bref qui se conclut par la prise de la drogue à effet amnésique. La Belle n'est réveillée à coup sûr lundi que si le Tirage amène 2, donc dimanche elle n'est pas certaine d'être réveillée. C'est pourquoi, quand elle se réveille lundi, elle acquiert indéniablement une information éternelle qu'elle n'avait pas dimanche, l'information R : « La Belle est réveillée lundi » ou « La belle est réveillée durant l'expérience ». Elle peut exprimer R par « Je suis réveillée… » : c'est sans conséquence puisqu'elle sait qu'elle est la Belle et que « je » n'est pas un indexical essentiel. Contrairement au problème n° 2, ici l'information nouvelle est, aux yeux d'un bayésien, pertinente pour l'issue du Tirage. Dimanche, la Belle accorde le crédit 1/2 à l'hypothèse H : « Le Tirage amène 1 ». Au réveil, elle révise cette croyance en conditionalisant sur R :

$$\begin{aligned} C_R(H) &= C(H \mid R) \\ &= C(R \mid H).C(H) / (C(R \mid H).C(H) + C(R \mid \neg H).C(\neg H)) \\ &= 1/2.1/2 / (1/2.1/2 + 1.1/2) \\ &= 1/3. \end{aligned}$$

1/3 est donc la réponse qu'une Belle bayésienne doit fournir lors de l'entretien avec les expérimentateurs. Le fréquentisme peut-il retrouver ce résultat ?

La question des expérimentateurs semble laisser au fréquentiste un choix : celui de la classe de référence. La probabilité que ce Tirage singulier ait amené 1 pourrait être directement inférée de la fréquence des issues-1 dans l'ensemble des issues de Tirages répétés un très grand nombre de fois, soit 1/2 comme dans les problèmes n$^{os}$ 1 et 2. Cependant, dans la situation de la Belle, le fréquentiste comprend qu'on ne lui demande pas n'importe quelle fréquence relative. Rappelons que, le dimanche, la probabilité bayésienne de l'issue 1, certes épistémique, est néanmoins fondée sur une propension ou une fréquence. Le fréquentiste, lui aussi, mélange un peu les genres : le choix de la classe de référence, pour être le plus rationnel, doit tenir compte des données disponibles, et la probabilité qui résulte de ce choix doit mesurer une disposition de la Belle à parier sur l'issue 1 au moment même où elle répond aux expérimentateurs. Aussi, si elle n'est pas insensible au versant épistémique des probabilités, une Belle fréquentiste comprend qu'on lui demande une fréquence qui pourrait devenir voire être interprétée comme *son* degré de croyance *sur le moment*. Elle sait qu'elle a plus de chances d'être réveillée si le Tirage a amené 2, elle doit donc croire en l'issue 1 faiblement, être prête à ne miser sur l'issue 1 qu'une petite somme en vue d'un gain supérieur. La fréquence adéquate est déterminée en restreignant la classe de référence : seuls doivent compter les Tirages suivis par un réveil lundi et donc par une occasion de s'interroger pendant l'expérience sur la probabilité de l'issue 1. Pour le voir autrement, cette fréquence est le rapport du nombre de réveils-1 (réveils au cours d'une expérience dont le Tirage a pour issue 1) sur le nombre total de réveils. Elle est clairement 1/3, puisqu'il y a deux fois plus de réveils-2 que de réveils-1 dans la série de réveils produite par la répétition (virtuelle) de l'expérience. 1/3, un sur *trois*, quantifie bien la disposition à parier *le lundi* : ce jour-là, un pari sur l'issue 1 n'est acceptable que si le sujet de l'expérience reçoit au moins *trois* fois sa mise en cas de succès, c'est-à-dire retrouve sa mise plus un bénéfice de deux fois la mise. Bien sûr, seul est concerné l'agent soumis au

protocole expérimental (qu'il connait) et susceptible d'être réveillé ou de rester endormi lundi ; un agent extérieur raisonne avec d'autres données.

Ainsi, 1/3 est la réponse adressée par une Belle fréquentiste aux savants qui la questionnent. Nous avons établi que c'est aussi la réponse bayésienne. Les tenants de deux interprétations concurrentes sont d'accord et, n'en doutons pas, cet accord, qu'il soit sur 1/2, 1/3 ou une autre valeur, demeure quelles que soient les probabilités placées dans les cases du tableau qui schématise une expérience d'un seul jour (lundi).

### 4.4. Problème n° 4

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| 1 | 1 | |

Cette fois-ci, l'unique question posée à la Belle réveillée au cours de l'expérience est : « Quelle est la probabilité que ce jour soit lundi ? »

Dimanche, la Belle est certaine que le Tirage amènera 1, certaine qu'elle sera réveillée une unique fois pour un entretien avec les scientifiques, mais elle ignore si ce réveil aura lieu lundi ou mardi. Clairement, les chances objectives d'un réveil lundi sont 1/2, que ce nombre soit la propension du tirage secondaire à désigner le lundi ou bien la fréquence des réveils-lundi parmi tous les réveils d'expériences virtuellement répétées. Un bayésien ajouterait que la probabilité épistémique est évidemment alignée sur ce résultat : la Belle doit croire au degré 1/2 qu'elle sera réveillée lundi. Ces observations et estimations sont nécessairement faites par un agent rationnel dimanche. Quand cet agent est réveillé au cours de l'expérience, il n'a donc pas le temps, pour ainsi dire, de ne pas être eccéitiste : certes, incapable de se repérer dans le temps au

jour près, il envisage un compartiment-lundi et un compartiment-mardi, il tente de probabiliser sa position temporelle, mais il est immédiatement invité au constat de l'équivalence de la croyance temporelle « Aujourd'hui est lundi » à la croyance éternelle « Le tirage secondaire a désigné lundi » ou « Un réveil a lieu lundi durant cette expérience ». Et comme il ne gagne aucune information susceptible de modifier les probabilités du dimanche, il doit répondre 1/2 à la question des expérimentateurs. Le bayésien et le fréquentiste seront d'accord, à n'en pas douter.

### 4.5. Problème n° 5

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| 1 | 1/2 | |

Comme dans l'exemple précédent, dans le cas où la Belle est réveillée au cours de l'expérience, on lui demande : « Quelle est la probabilité que ce jour soit lundi ? »

Le raisonnement est assez similaire à celui qui conclut 1/2 dans le problème précédent, mais une petite difficulté se présente ici. La Belle n'est pas certaine, dimanche, qu'elle sera réveillée dans la période lundi-mardi. Dimanche, la probabilité d'un réveil lundi n'est donc plus 1/2 mais 1/4 : c'est évidemment le produit de la probabilité d'un réveil dans la période lundi-mardi (1/2) et de la probabilité d'un réveil lundi plutôt que mardi sachant qu'il y a réveil dans la période lundi-mardi (1/2). Lorsque la Belle se réveille pendant l'expérience, elle apprend par là même qu'elle est réveillée dans la période lundi-mardi et, de son point de vue, qu'elle soit bayésienne ou fréquentiste, la probabilité d'un réveil lundi redevient 1/2 : la Belle bayésienne augmente de 1/4 à 1/2 un degré de croyance par conditionalisation sur l'information nouvelle, tandis que la Belle

fréquentiste considère une classe de référence plus étroite dans laquelle la fréquence des réveils-lundi est 1/2 (les raisons de ce changement de classe sont précisées dans l'analyse du problème nº 3). Réveillée durant l'expérience, la Belle attribue donc la probabilité 1/2 à « Un réveil a lieu lundi durant cette expérience », qui, dans le contexte de l'expérience, est l'équivalent éternel de la croyance temporelle « Aujourd'hui est lundi ». Ainsi, 1/2 est la réponse à donner aux expérimentateurs. Cette valeur n'est pas étonnante : on ne voit pas pour quelle raison la Belle préférerait lundi à mardi, ou mardi à lundi. Nous avons le sentiment que le principe d'indifférence s'étend aux possibilités temporelles. Le problème suivant montre mieux cette extension.

### 4.6. Problème nº 6

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| 1 | 1 | 1 |

La Belle est ici nécessairement réveillée deux fois pendant l'expérience. On lui demande à chaque fois : « Quelle est la probabilité que ce jour soit lundi ? »

Avec ce problème compartimenté, des subtilités surviennent. Dimanche, la Belle est certaine qu'elle sera réveillée pendant l'expérience, certaine qu'elle sera réveillée lundi, certaine aussi qu'elle sera réveillée mardi. Elle conserve ces certitudes au réveil, qu'il soit le réveil-lundi ou le réveil-mardi ; pourtant elle est alors dans l'incertitude en ce qui concerne sa position temporelle. Dans le problème nº 4, le protocole l'autorisait à penser : « Aujourd'hui est lundi si et seulement si un réveil a lieu lundi durant l'expérience ». Le protocole présent détruit cette équivalence. Nous avons néanmoins le sentiment qu'une probabilité peut être rationnellement

associée à l'hypothèse temporelle « Aujourd'hui est lundi » et que cette probabilité est 1/2. Au moins trois arguments mènent à cette conclusion : un argument bayésien temporaliste, très représenté dans la littérature, est concurrencé par un argument bayésien eccéitiste et par un argument fréquentiste « amélioré ».

Les premiers analystes de la Belle au bois dormant, parmi lesquels David Lewis, ont sur certains points de leur discours, voire à chaque étape de celui-ci, une vision « temporaliste » des croyances auto-localisantes. Dans les commentaires de Darren Bradley, le temporalisme est l'idée selon laquelle une croyance, comme emportée dans le cours des choses, a un contenu dont la valeur de vérité est changeante. Nous désignons par temporaliste un analyste qui s'oppose à la réduction eccéitiste à une croyance éternelle, au moins face aux situations où, comme ici, un sujet se trouve en même lieu à plusieurs moments différents, mais pas nécessairement face aux situations où plusieurs sujets se trouvent en des lieux différents au même moment (le cas des deux dieux de Lewis). Nous avons bien affaire ici à deux centres ou deux parties temporelles d'un être continu, où plutôt dont la continuité est brisée par l'amnésie : dans un même monde un même sujet s'éveille deux fois qui ne lui en paraissent qu'une. Une fois, le jour est lundi *et pas encore* mardi, l'autre fois le jour est mardi *et plus du tout* lundi. Comment croire que ce jour est lundi, comment croire que la vérité de « Aujourd'hui est lundi » ne s'est pas mutée en fausseté avec le temps dont le flux échappe aux amnésiques ? Réponse du temporaliste : en croyant partiellement que c'est lundi, avec la même intensité que pour la croyance adverse « Aujourd'hui est mardi ». Une valeur de vérité qui se change en son autre, c'est justement ce qui fait d'une proposition une hypothèse, une possibilité, que ce changement ait lieu quand on envisage un à un des mondes ou parce que le temps fuit. Il faut suivre Lewis et penser que le principe d'indifférence s'applique aussi quand on a affaire à des centres colocalisés, à partir du moment où est

satisfaite la condition : aucune raison de privilégier une des possibilités. Le principe fut peut-être pensé pour équilibrer des hypothèses considérées comme éternelles, mais rien au fond ne l'empêche de s'appliquer pour un champ plus vaste d'hypothèses. C'est pourquoi la Belle doit répondre 1/2 à la question des scientifiques.

L'eccéitisme aussi respecte le principe d'indifférence et répond 1/2, mais l'extension du principe aux possibilités temporelles est expliquée par l'équivalence à des possibilités éternelles. Pendant le réveil du lundi se produisent dans le monde des événements uniques. La Belle, certes, n'est pas omnisciente, mais on peut se demander à quel point elle est ignorante. Réveillée au cours de l'expérience, est-elle coupée du monde à tel point que, par exemple, la mouche qui se pose sur sa main gauche dix minutes après sa sortie du sommeil est cachée à ses sens ? Si la réponse est non, « Aujourd'hui est lundi » est pour la Belle équivalente à « Cette mouche se pose sur ma main gauche lundi dix minutes après mon réveil » ; de même, « Aujourd'hui est mardi » équivaut à « Cette mouche se pose sur ma main gauche mardi dix minutes après mon réveil ». Si la réponse est oui, la Belle peut quand même penser que son présent réveil n'est pas totalement identique à l'autre réveil de l'expérience, le réveil du mardi ne peut pas être « le retour du même » réveil juste parce que lundi une drogue efface chez le sujet de l'expérience tout souvenir du réveil. Par un grand effort d'introspection, la Belle peut parvenir à saisir l'unicité de l'instant vécu, son très singulier état d'esprit agité par divers stimulants extérieurs. Ce vécu particulier a lieu lundi ou bien mardi, pas les deux jours. D'une façon ou d'une autre, la Belle peut ramener une croyance temporelle à une croyance éternelle. Selon un eccéitiste, le principe d'indifférence s'applique lorsqu'on suppose que les expérimentateurs sont suffisamment puissants pour parvenir à cacher à la Belle tout indice qui la ferait pencher vers lundi plutôt que vers mardi ou inversement : la Belle a alors un vécu unique sans aucune pertinence pour sa position temporelle, mais ce vécu

suffit à lui faire envisager deux hypothèses éternelles exclusives et conjointement exhaustives, équivalentes aux hypothèses temporelles et rendues équiprobables par indifférence.

Le fréquentisme, lui aussi, fait face à un problème sérieux : si l'expérience était répétée un très grand nombre de fois, les réveils-lundi et les réveils-mardi ne surviendraient plus au hasard comme dans le problème n° 4. Il y aurait nécessairement une alternance des réveils : un réveil-lundi, puis un réveil-mardi, puis un réveil-lundi, puis un réveil-mardi, et ainsi de suite. Il est donc impossible de parler d'une série *aléatoire* de réveils, impossible d'identifier la fréquence des réveils-pile dans cette série à une probabilité[122]. En outre, on ne voit plus quel événement, effet de quelque dispositif aléatoire, est désigné par « Aujourd'hui est lundi ». Que signifie alors, pour la Belle, la probabilité que ce jour soit lundi ? L'hypothèse SSSA[123] de Bostrom et les diverses tentatives pour justifier l'analogie de la Belle au bois dormant avec un problème d'urne montrent que des chercheurs qui ont tendance à interpréter statistiquement la probabilité invitent à considérer le réveil de la Belle en un jour particulier comme le résultat d'un processus aléatoire subtil dont le tirage au sort est une métaphore. La Belle en tant qu'individu continu se réveille lundi puis mardi, mais en tant qu'individu temporel elle se réveille lundi *ou* mardi ; si elle associe une probabilité non extrême à « Aujourd'hui est lundi », c'est qu'elle considère son moment présent comme le résultat ignoré d'une

[122] Delabre et Gerville-Réache (2015) évoquent particulièrement ce point. Une statistique est habituellement une somme de *n* variables aléatoires de Bernouilli indépendantes et identiquement distribuées divisée par *n* : cette statistique seule converge (quand *n* tend vers l'infini) vers une probabilité au sens de la théorie fréquentiste traditionnelle. Il n'est pas interdit de penser des probabilités statistiques différentes, mais surviennent alors un grand nombre de questions qui doivent trouver réponses.

[123] Dans le chapitre 2, nous avons ainsi énoncé la *Strong Self-Sampling Assumption* : « Un observateur devrait raisonner comme si son moment présent était un échantillon aléatoire de l'ensemble de tous ses moments dans sa classe de référence. »

élection d'un moment parmi les moments dénombrables contenus dans une « urne métaphysique ». L'urne, ici, contient un lundi et un mardi ; un des deux, comme tiré au sort, devient le moment présent de la Belle. Le tirage est équilibré en l'absence de données qui permettraient d'associer à lundi et à mardi des poids différents. La probabilité 1/2 qu'il convient d'attribuer à « Aujourd'hui est lundi » n'est pas physique mais elle est ontique, et elle peut être aussi statistique : de toute urne matérielle contenant deux objets A et B qu'on ne peut distinguer qu'hors de l'urne, on pourrait extraire A une fois sur deux tirages (avec remise) si l'on procédait à un très grand nombre de tirages ; l'urne métaphysique est analogue.

### 4.7. Problème n° 7

| Jour / Tirage | **Lundi** | **Mardi** | **Mercredi** |
|---|---|---|---|
| 1 | 1 | 1 | 1 |

Dimanche, les expérimentateurs conviennent avec la Belle qu'ils cacheront une rose sous son oreiller dans la nuit de mardi à mercredi. Lors de chacun des trois réveils, ils lui demanderont : « Quelle est la probabilité qu'une rose se trouve sous votre oreiller ? »

Cette question semble bien attendre la même réponse que la question « Quelle est la probabilité que ce jour soit mercredi ? » Nous dirions de prime abord que la différence entre ce problème et le précédent est mince et que l'analyse précédente, à quelques détails près, est opportune ici aussi. C'est vrai. Et la Belle, bayésienne ou fréquentiste, devrait répondre 1/3, un sur trois puisqu'il y a non plus deux mais trois compartiments. Néanmoins, les réactions de quelques chercheurs avec qui j'ai eu l'occasion de discuter, notamment le statisticien Léo Gerville-Réache et le philosophe Paul

Franceschi, entretiennent un doute. Difficile à défendre, la réponse 1/2 ne doit pourtant pas être préjugée irrationnelle et écartée trop hâtivement.

D'abord, la nature de « Une rose se trouve sous mon oreiller » n'est pas claire. Est-ce une proposition auto-localisante ou une proposition qu'on rend équivalente à une proposition auto-localisante ? Elle décrit un état du monde mais pas à un temps absolu, elle est essentiellement indexicale puisque l'indexical « maintenant » est sous-entendu et que la Belle ne saurait le remplacer par une donnée temporelle univoque. « Une rose se trouve sous mon oreiller » ne parait être qu'un raccourci pour « Je suis à un moment où une rose se trouve sous mon oreiller » ou « Aujourd'hui est un jour où une rose se trouve sous mon oreiller », qui localise le sujet dans le temps bien que cette localisation soit explicitée laborieusement, par périphrase.

Ensuite, imaginons une variante où la Belle ne sait pas durant quelle nuit une rose est cachée sous son oreiller par les scientifiques. La réponse à la question de l'entretien sera 1/2 en vertu du principe d'indifférence : la rose est là, ou bien elle n'est pas là, et rien ne permet de préférer une des deux options. Que signifie la réponse 1/3 lorsque la Belle sait que la rose est cachée dans la nuit de mardi à mercredi ? Elle signifie qu'il y a dans ce cas une raison de préférer la possibilité de l'absence de la rose : seul un réveil sur les trois que comporte l'expérience est un réveil agrémenté par le cadeau sous l'oreiller et, pour la Belle, se trouver dans un réveil-lundi, un réveil-mardi, un réveil-mercredi sont trois hypothèses équiprobables… en vertu du principe d'indifférence (si elle est sensible au versant épistémique de la probabilité). Le souci dans ce raisonnement est qu'on applique le principe d'indifférence aux trois positions temporelles pour en déduire la probabilité de la présence de la rose, sans se demander s'il ne serait pas plus légitime d'appliquer le principe aux deux alternatives présence/absence de la rose avant d'en déduire éventuellement la probabilité d'être mercredi. Si cela doit mener à une probabilité 1/2

contrintuitive, rappelons-nous que probabiliser les moments d'un agent est quelque chose d'assez nouveau, si ce n'est d'assez suspect.

Il se peut que les scénarios de multiplication des moments possibles gênent plus particulièrement les demistes, les pro-1/2 du problème original de la Belle au bois dormant. Observons deux petites énigmes jumelles que j'ai proposées à Paul Franceschi lorsqu'il était encore demiste.

Un beau matin, vous vous réveillez, mais pas chez vous : vous êtes seul dans une chambre close inconnue. On vous apprend par un haut-parleur qu'on vous a kidnappé dans votre sommeil il y a de cela plusieurs mois, pendant lesquels on vous a maintenu inconscient et en bonne santé. Vous êtes en ce moment même le cobaye d'une expérience. On vous demande la probabilité que cette semaine soit une semaine paire du calendrier. On vous met en garde : si cette semaine est effectivement paire, on ne vous pose la question qu'une fois, autrement dit juste maintenant, mais si ce n'est pas le cas, on vous la pose deux fois. À chaque fois on vous explique ces règles, et après l'entretien on vous drogue et vous vous endormez en oubliant votre réponse et tout ce qui s'est passé pendant vos quelques instants de conscience, ainsi rien ne peut vous indiquer si la question vous a déjà été posée et, bien sûr, rien ne peut vous indiquer la parité de cette semaine. Qu'allez-vous répondre ? En tout cas, gageons qu'un tiériste répondra 1/3, alors qu'un demiste répondra 1/2, comme s'ils étaient engagés dans l'expérience originale de la Belle.

Énonçons une variante. Vous vous réveillez dans une chambre inconnue. On vous apprend qu'on vous a kidnappé et que vous êtes en ce moment même le cobaye d'une expérience. On a décidé de vous plonger dans un sommeil amnésique de deux cents semaines, on vous réveille une heure tous les lundis et un mardi sur deux, le mardi des semaines impaires, pour vous expliquer ces règles et vous demander la probabilité que cette semaine soit une semaine paire. Vous ne pouvez dater ni ce jour ni même

cette semaine. N'est-ce pas la réponse 1/3 qui vous séduit le plus ? Dans le courriel qu'il m'adressait, Franceschi comprenait que l'on puisse intuitionner la probabilité 1/3 même en étant demiste, il savait que Nick Bostrom, par exemple, répondrait 1/3[124], et pourtant il restait tenté par la réponse 1/2. Il pouvait s'aider de l'énigme jumelle : qui se trouve dans la situation exposée ici, et ainsi se réveille trois cents fois sur deux cents semaines, se trouve apparemment aussi dans la situation vue juste au-dessus, il est réveillé pendant une semaine dont il ignore la parité. La pensée de la répétition des semaines à un ou deux réveils perturbe et divise les demistes. Tant que la solution demiste de la Belle au bois dormant fait partie du débat, ce problème n° 7 à trois compartiments (et une rose) gardera lui aussi une part de mystère malgré toute la force de l'intuition de la réponse 1/3.

### 4.8. Problème n° 8

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **1** | 1 | 1 |
| **2** | 1 | 1 |

Les expérimentateurs ont une seule question à poser à la Belle chaque fois qu'elle se réveille dans l'expérience : « Quelle est la probabilité que le tirage ait amené 1 et que ce jour soit lundi ? »

[124] Bostrom (2007) défend la réponse 1/2 pour le problème original de la Belle ; l'argumentation, qui emploie des variantes où l'expérience est étalée sur de très nombreuses semaines, montre que Bostrom répondrait sans aucun doute 1/2 à la première de nos deux énigmes jumelles, 1/3 à l'autre.

On demande ici de peser une hypothèse temporelle mixte ou auto-localisante au sens large, une possibilité centrée qui dit aussi quelque chose du monde : « Le Tirage amène 1 *et* on est lundi ». La réponse, 1/4 évidemment, est tirée des analyses des problèmes précédents. Pour une Belle bayésienne, « Le Tirage amène 1 » a pour probabilité (épistémique) 1/2 en vertu du principe principal et du principe d'inertie doxastique, tandis que « On est lundi » a pour probabilité 1/2 en vertu du principe d'indifférence ; les réveils sont indépendants du Tirage, et le calcul des probabilités se moque de la provenance des deux résultats 1/2, donc le produit 1/2.1/2 est la probabilité de l'hypothèse mixte. Pour une Belle fréquentiste, « Le tirage amène 1 » a pour probabilité (ontique) 1/2 d'après l'examen des effets d'un dispositif aléatoire physique, tandis que « On est lundi » a pour probabilité 1/2 d'après l'examen des effets d'un processus aléatoire métaphysique, indépendants des effets du premier dispositif ; le calcul des probabilités donne là encore 1/2.1/2.

Finalement, la Belle associe plutôt aisément la probabilité 1/4 à chacun des quatre compartiments qu'elle envisage au réveil. Il suffirait qu'elle apprenne lors d'un entretien avec les scientifiques que, par exemple, $<<$ Belle, mardi $>$, 1 $>$ n'est assurément pas actuel (est une impossibilité) pour se retrouver dans la situation décrite dans le problème original, et ainsi se retrouver en grande difficulté pour probabiliser les compartiments restants. Certes, à sa place nous aurions tendance à partager équitablement l'ancienne probabilité 1/4 de la possibilité éliminée sur les trois autres. Pourtant, souvenons-nous du problème de la machine à dupliquer, vu dans le chapitre 2, et craignons l'existence d'une critique efficace de ce partage équitable.

### 4.9. Problème n° 9

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| 1 | 1/2 | 1/2 |

Les expérimentateurs ont deux questions à poser à la Belle en cas de réveil : « Quelle est la probabilité que ce jour soit lundi ? » et « Quelle est la probabilité que vous soyez réveillée deux fois pendant l'expérience ? »

Les analyses des problèmes précédents nous incitent à penser que la réponse à la première question est 1/2 et que fréquentisme et bayésianisme s'entendent sur ce résultat. Les probabilités d'un réveil lundi et d'un réveil mardi sont identiques. Comment fréquence et crédit pourraient différer de 1/2 ? Répondre à la seconde question est en revanche extrêmement difficile. D'ailleurs, personne de sérieux n'est capable aujourd'hui de donner une réponse qui engage la communauté scientifique et philosophique. Comment le savons-nous ? Parce que ce problème présente les mêmes caractéristiques déconcertantes que le problème original de la Belle au bois dormant. Montrons-le.

Le générateur G2 est sollicité à deux reprises dans ce scénario : une fois pour déterminer équitablement si on doit réveiller la Belle lundi ou bien si on doit la laisser dormir, une fois pour déterminer équitablement si on doit la réveiller mardi ou la laisser dormir. Du point de vue des expérimentateurs, commencer une expérience avec la Belle, c'est nécessairement amener une de ces quatre possibilités équiprobables : soit la Belle n'est réveillée ni lundi ni mardi, soit elle est réveillée lundi seulement, soit elle est réveillée mardi seulement, soit elle est réveillée lundi et à nouveau mardi. Ainsi notre protocole ressemble à celui-ci :

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **1** | 0 | |
| **2** | 1 | 0 |
| **3** | 0 | 1 |
| **4** | 1 | 1 |

La seule différence est que dans cette variante on ne consulte plus G2 mais G1 pour savoir quand réveiller le sujet, et on le fait dès le dimanche soir. Les réponses de la Belle ne peuvent pas être modifiées par ce changement. Nous voyons plus clairement à présent pourquoi il est si difficile de répondre à la seconde question des scientifiques. La Belle ignore dimanche combien de fois elle sera réveillée, comme dans le problème original. Lorsqu'elle est réveillée dans l'expérience, elle apprend qu'elle n'est pas dans une expérience où elle n'est réveillée ni lundi ni mardi. Une conditionalisation sur cette information lui permettrait de réviser sa croyance en « On me réveille deux fois durant l'expérience », dont le degré passerait de 1/4 à 1/3. Ce nombre, un sur trois, est prévisible, il se lit dans le tableau : il n'y a plus que trois scénarios possibles une fois qu'est éliminée la possibilité d'un sommeil ininterrompu dans la période lundi-mardi. Pourtant, si l'expérience était répétée de très nombreuses fois, on assisterait à une série de réveils qui comporterait, comme le montre aussi le tableau, autant de réveils uniques au sein de leur expérience que de réveils dans une expérience où ont lieu deux réveils. Ce n'est donc pas à 1/3 mais à 1/2 que les analyses fréquentistes précédentes devraient nous mener.

Fortuitement, l'hypothèse temporelle « Aujourd'hui est lundi » est plus facile à probabiliser que l'éternelle, « On réveille la Belle deux fois durant l'expérience ». Cela se voit aisément dans notre tout dernier tableau

issu d'une opération, pas toujours réalisable, de redécoupage des compartiments schématisés par le tableau d'origine afin de ne montrer que des cases marquées par une probabilité extrême, 0 ou 1. La disparité entre les lignes 2 et 3 (un réveil donc un compartiment) et la ligne 4 (deux réveils donc deux compartiments) est annonciatrice d'un problème sérieux. Mais les deux colonnes présentent le même nombre de cases marquées 1 : tant que la Belle, lors d'un entretien avec les expérimentateurs, n'est pas informée ni poussée à un mouvement des probabilités qu'elle assigne aux divers compartiments, elle n'a pas de difficulté à probabiliser une hypothèse auto-localisante au sens de Perry, telle que « Aujourd'hui est lundi » qui purement localise dans le temps, c'est-à-dire ne dit pas, de surcroit, comment est le monde.

# Chapitre 4

# La Belle au bois dormant

Nous nous sommes familiarisés avec des problèmes d'auto-localisation compartimentés. Ils n'étaient pas tous faciles à résoudre et certains présentent encore des zones d'ombre, mais ce n'est rien comparé à la Belle au bois dormant. Nous allons rester assez longtemps sur cette énigme, la littérature est vaste et les solutions apportées forment un parc d'attractions merveilleux pour amateurs de paradoxes, où il ne faut pourtant jamais trouver fantaisiste un argument qui a l'air de soutenir qu'un chat a six pattes parce qu'il est analogue à un insecte.

Le paradoxe peut être présenté de deux manières, que nous allons détailler. Nous comprendrons d'abord qu'il crée un désaccord imprévu entre fréquentistes et bayésiens, qui s'opposent notamment sur deux probabilités, 1/3 chez les premiers, 1/2 chez les seconds, scission d'autant plus inquiétante que de tout côté des raisonnements apparemment valides utilisent des principes d'alignement. Nous donnerons ensuite les différents types de résolution trouvables dans la littérature. Les « tiéristes », nombreux mais divisés, se querellent avec les « demistes » et les « doubles demistes » ; les rares « désambiguïseurs » pimentent encore le débat. La place de la double interprétation de la probabilité dans chaque tentative de

résolution sera analysée. En parcourant les textes des chercheurs, nous observerons chez certains une direction qui les dépasse, une tendance à réparer la scission paradoxale en faisant de 1/2 une probabilité ontique fréquentielle et plus seulement épistémique, ou bien en faisant de 1/3 une probabilité épistémique bayésienne et plus seulement ontique.

## 1. La Belle au bois dormant, problème compartimenté

Le protocole peut être schématisé très simplement. Modifions un détail dans nos habitudes afin de coïncider avec l'énoncé canonique : remplaçons le générateur G1 par une pièce de monnaie équilibrée.

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **Face** | 1 | 0 |
| **Pile** | 1 | 1 |

La question canonique posée par les expérimentateurs lors d'un réveil est : « À quel degré devez-vous croire que la pièce est tombée sur face ? » Est donc précisément demandé à la Belle un degré de croyance, non une probabilité laissée à sa libre interprétation. Mais le fréquentisme ne cesse pas pour autant d'avoir son mot à dire.

### 1.1. Première appréciation du paradoxe

Une manière de goûter le paradoxe est de considérer les premiers arguments que la Belle pourrait soutenir après un réveil dans l'expérience. Le raisonnement bayésien le plus intuitif, en accord avec nos réflexions du chapitre 3, conduit à une position demiste, tandis que le raisonnement

fréquentiste le plus immédiat et le plus conforme à nos dernières réflexions conclut la réponse tiériste.

Le raisonnement bayésien travaille une probabilité épistémique introduite dès le départ par alignement sur des chances objectives. La propension de la pièce à tomber du côté face dimanche soir est 1/2 et la Belle, juste avant de s'engager dans l'expérience, ne reçoit pas d'information inadmissible[125] qui disqualifierait cette simple donnée, c'est pourquoi la Belle doit respecter le principe principal et son degré initial de croyance en face doit être 1/2. Quand elle est réveillée dans l'expérience, événement qui n'est pas pour elle une surprise, elle maintient ce degré de croyance parce qu'elle respecte le principe d'inertie doxastique : pas d'information nouvelle pertinente pour l'issue pile/face, pas de révision de la croyance en face. Elle doit donc répondre 1/2 à la question des expérimentateurs. Ce raisonnement est à peu de choses près celui qui est tenu par David Lewis dans son unique article sur la Belle au bois dormant ; il détaille le court raisonnement imprécis formulé par Adam Elga lorsque ce dernier introduisit le paradoxe[126]. Remarquons que ce degré 1/2 d'une croyance éternelle est évalué sans que soient envisagés des compartiments ou des possibilités centrées. Si maintenant la Belle admet qu'un principe d'indifférence la contraint à rendre équiprobables « La pièce tombe sur pile et on est lundi » et « La pièce tombe sur pile et on est mardi », la probabilité de chacune de ces deux hypothèses doit être la moitié de 1/2, soit 1/4.

Le raisonnement fréquentiste opte pour une lecture ontologique fictionnelle de la probabilité cherchée, en présupposant que celle-ci ne varie pas avec le nombre d'expériences réellement programmées : une seule expérience ou mille expériences ne changeront rien. La Belle imagine

[125] Inadmissible au sens lewisien, vu au chapitre 2.

[126] Elga (2000) ; D. Lewis (2001).

donc que l'expérience dans laquelle elle est engagée est répétée de très nombreuses fois et elle considère une très longue série de réveils. Les réveils sont *a priori* les objets à compter pour établir une fréquence, étant donné que deux réveils peuvent avoir lieu par expérience et que dans des cas semblables seul l'objet réveil correspond à un compartiment envisagé par la Belle, seule la focalisation sur les réveils élémentaires permet de savoir si la position temporelle de la Belle influence la probabilité de l'issue du Tirage. La Belle constate qu'un réveil sur trois est un réveil-face. La statistique 1/3 se dégage. Le souci est peut-être la nature semi-aléatoire de la série de réveils, dans laquelle il est impossible qu'un réveil-pile soit isolé, dans laquelle nécessairement un réveil-pile en côtoie un autre, voire plusieurs autres. Il faut alors recourir à l'urne métaphysique : la Belle doit considérer qu'elle a deux fois plus de chances de vivre maintenant un réveil-pile, comme si ce moment était tiré d'une urne contenant deux fois plus d'objets-pile que d'objets-face. La Belle est donc presque prête à répondre 1/3 aux expérimentateurs, il faut simplement qu'elle croit en face au degré 1/3 après avoir calqué cette probabilité épistémique sur la probabilité ontique qu'elle a évaluée, en suivant le principe de Miller ou une de ses variantes. Notons que, pour un agent fréquentiste non engagé dans l'expérience, et notamment pour la Belle du dimanche, la probabilité de face est évidemment 1/2, soit la fréquence des faces parmi les résultats de lancers répétés de la pièce équilibrée. Le protocole contraint le sujet de l'expérience à prendre en compte d'autres objets et d'autres proportions, d'où une estimation différente de la probabilité.

Le paradoxe est d'abord cette surprenante contradiction de deux conclusions : nous refusons la coexistence de deux réponses, 1/2 et 1/3, pourtant appuyées par des observations et des inférences qui nous sont maintenant familières. Un écart peu ordinaire entre deux probabilités obtenues très différemment s'est creusé. Nous pensons que des agents rationnels doivent craindre des divergences inexpliquées entre degrés de

croyance et chances objectives. Pourtant, l'argument bayésien manifeste la volonté de respecter un principe d'alignement. Il en est de même de l'argument fréquentiste. Mais l'alignement n'a pas du tout lieu à la même étape. Le raisonnement bayésien est introduit par un alignement entre une propension et un degré de croyance ; la croyance en face progresse ensuite sans attache physique, ne se compare plus à des propensions ni à des fréquences, et le raisonnement conclut une probabilité 1/2 qu'on doit qualifier d'épistémique malgré son « hybridation » passée avec une probabilité ontique. Le raisonnement fréquentiste est quant à lui la recherche de la juste fréquence de certains événements, de certaines propriétés, dans la bonne classe de référence, et l'alignement avec un degré de croyance n'a lieu qu'à la fin, parce que ce degré est exigé par les expérimentateurs. Ainsi ce raisonnement s'attarde dans le monde (bien qu'il soit le monde d'une simulation mentale, le monde de la répétition de l'expérience) et délaisse longtemps le visage épistémique de la probabilité, ce qui fait de son résultat 1/3 une probabilité plus ontique qu'épistémique.

### 1.2. Deuxième appréciation du paradoxe

Il est temps d'introduire ce que nous appellerons la question subsidiaire du lundi, un complément à l'énoncé du problème, nécessaire à beaucoup de chercheurs pour développer une analyse dans de bonnes conditions. Lorsque c'est lundi et que la Belle a répondu à la question principale des expérimentateurs, ceux-ci lui apprennent que c'est lundi avant de lui reposer la question : « À quel degré devez-vous croire que la pièce est tombée sur face ? » Rappelons que dimanche, juste avant l'expérience, la Belle connait tout le protocole, donc sait que lundi la question lui sera posée une première fois, alors qu'elle sera encore incapable de se repérer dans le temps, puis une seconde fois, quand elle sera sûre que son réveil est un réveil-lundi. Notre problème, c'est toujours de trouver ce que la Belle doit répondre à la première question du lundi, qui

est aussi l'unique question du mardi éventuel (si pile), mais à présent nous cherchons aussi ce qu'elle doit répondre à la seconde question du lundi.

Cette question subsidiaire va nous permettre de présenter une autre manière d'apercevoir le paradoxe, plus évoluée mais complémentaire de la première. Notre objectif est d'ouvrir assez naturellement la voie vers les grands types de résolutions du problème rencontrés dans la littérature. Nous n'allons pas exposer, comme cela se fait souvent, les étapes d'un raisonnement apparemment valide menant pourtant à une conclusion inacceptable, puis chercher, étape après étape, où l'erreur a pu se glisser. Non. Nous allons procéder avec plus d'originalité, mais avec efficacité.

Avant de continuer, quelques conventions en très petit nombre s'imposent. Les questions de l'énoncé exigent des probabilités épistémiques, plus exactement des degrés rationnels de croyance. Les trois mesures du crédit que la Belle accorde à ses croyances le dimanche, puis au moment où elle se réveille dans l'expérience, et enfin après l'annonce « on est lundi » seront notées respectivement : $C_{dim}(\cdot)$, $C_{rév}(\cdot)$ et $C_{lun}(\cdot)$. Ces distributions satisferont les axiomes de Kolmogorov. Cinq hypothèses, dont trois sont centrées et plus exactement temporelles (le sujet qui les exprime cherche à se localiser dans le temps) intéresseront particulièrement la Belle :

F : « La pièce tombe sur face »

P : « La pièce tombe sur pile »

FL : « La pièce tombe sur face et on est lundi » (« Je suis dans un réveil-face-lundi »)

PL : « La pièce tombe sur pile et on est lundi » (« Je suis dans un réveil-pile-lundi »)

PM : « La pièce tombe sur pile et on est mardi » (« Je suis dans un réveil-pile-mardi »)

Remarquons que, pour tout sujet *engagé dans l'expérience*, FL, PL et PM sont exclusives et conjointement exhaustives (de même F et P) : la somme des probabilités associées à ces trois hypothèses est égale à 1, même lorsque PM est disqualifiée par l'annonce de la date, lundi. Voici d'autres égalités qui ne sont jamais remises en cause dans les débats :

(i) $C_{dim}(F) = C_{dim}(P) = 1/2$, valeur obtenue par alignement, le dimanche, sur la propension de la pièce équilibrée à tomber sur face ou sur pile (principe principal de Lewis).

(ii) $C_{rév}(PL) = C_{rév}(PM)$, en vertu d'un principe d'indifférence étendu aux croyances localisantes, plus exactement aux mondes centrés colocalisés[127].

(iii) $C_{lun}(FL) = C_{lun}(F)$, en vertu d'un principe d'équiprobabilité de propositions logiquement équivalentes étendu à des propositions auto-localisantes qui ne peuvent être équivalentes à d'autres propositions que dans un certain contexte (ici, FL et F ne sont équivalentes pour la Belle que durant l'expérience, où un réveil-face ne peut avoir lieu qu'un lundi).

Posons maintenant quatre résultats, ($R_1$), ($R_2$), ($R_3$) et ($R_4$), issus d'une première analyse qui les montre comme difficilement contestables :

($R_1$) $C_{rév}(F) = C_{dim}(F)$, par inertie doxastique entre dimanche et le réveil.

($R_2$) $C_{lun}(F) = 1/2$, par alignement sur les chances objectives le lundi, en considérant que le résultat du Tirage n'a aucune conséquence sur les

[127] L'analyse du problème n° 6 dans le chapitre précédent essaie de justifier ce principe d'indifférence étendu.

événements du lundi[128] (en apprenant qu'elle est réveillée lundi, événement qui devait assurément arriver, la Belle ne fait que récupérer ce que le protocole lui a temporairement pris, c'est-à-dire la faculté de se situer dans le temps au jour près ; plus rien d'important ne la distingue d'un individu extérieur à l'expérience qui aligne son degré de croyance en face sur la propension de la pièce à tomber sur face, tant que rien ne lui fait préférer un côté de la pièce plutôt que l'autre).

($R_3$) $C_{lun}(FL) > C_{rév}(FL)$, par conditionalisation et vérification du « pouvoir confirmant » de l'information « on est lundi » = $FL \vee PL$[129].

($R_4$) $C_{rév}(FL) = C_{rév}(F)$, en vertu du principe étendu d'équiprobabilité de propositions logiquement équivalentes (voir l'égalité (iii) plus haut).

Le souci est qu'il est impossible de tenir pour vraies ces quatre relations en même temps : à l'évidence, on conclurait à partir d'elles des contradictions. Par exemple, de (i), ($R_1$) et ($R_2$), on déduit $C_{rév}(F) = C_{lun}(F)$, alors que de (iii), ($R_3$) et ($R_4$) on déduit $C_{rév}(F) < C_{lun}(F)$ : la contradiction patente, si l'on veut être démonstratif à l'extrême, conduit à l'aberration $C_{rév}(F) < C_{rév}(F)$ ! Un de nos raisonnements familiers utilisant l'inertie doxastique, le principe principal, la conditionalisation ou l'équivalence logique serait-il erroné ? Voilà qui nous permet d'apprécier le paradoxe sous un angle qui, *pour l'instant*, met de côté la rivalité ontique-

[128] En exposant cette démonstration, Adam Elga est encore plus prudent : il anticipe la critique selon laquelle une pièce qui a déjà été lancée n'a pas de propension non extrême de tomber sur face. Aussi, il imagine une expérience de la Belle dans laquelle ce Tirage n'a lieu que dans la nuit de lundi à mardi : en effet, ce n'est que dans la journée de mardi que l'issue pile/face a des conséquences, donc cette modification légère du scénario ne peut pas modifier notre analyse du problème.

[129] Par conditionalisation et application du théorème de Bayes, on trouve que :

$$C_{lun}(FL) = C_{rév}(FL \mid FL \vee PL)$$
$$= C_{rév}(FL \vee PL \mid FL).C_{rév}(FL) / C_{rév}(FL \vee PL)$$
$$= C_{rév}(FL) / C_{rév}(FL \vee PL).$$

En constatant que $0 < C_{rév}(FL \vee PL) < 1$, on déduit l'inégalité cherchée.

épistémique, mais qui a un avantage : si nous considérons qu'un des quatre résultats précités est faux, nous entrevoyons des pistes de résolution. Il se trouve que les chercheurs se concentrent sur la critique de l'un ou l'autre de ces quatre résultats et n'explorent pas d'autres pistes. Les quatre couples de réponses apportées à la question principale et à la question subsidiaire des expérimentateurs suivent les refus ou corrections de ($R_1$), de ($R_2$), de ($R_3$) ou de ($R_4$). Les chercheurs se répartissent ainsi :

| Type de résolution | Résultat rejeté | Réponse à la question principale | Réponse à la question subsidiaire |
|---|---|---|---|
| tiérisme | ($R_1$) | 1/3 | 1/2 |
| demisme | ($R_2$) | 1/2 | 2/3 |
| double demisme | ($R_3$) | 1/2 | 1/2 |
| désambiguïsation | ($R_4$) | 1/2 et 1/3 | 1/2 |

## 2. Le tiérisme

Commençons notre exploration par le refus du seul résultat ($R_1$) $C_{rév}(F) = C_{dim}(F)$, remplacé par $C_{rév}(F) = 1/3$, positionnement communément appelé tiérisme dans la littérature en français : effectivement, même s'ils répondent 1/2 à la question subsidiaire du lundi, les tiéristes répondent 1/3 à la question principale de l'entretien, et non 1/2 comme la plupart de leurs adversaires. Cette correction de ($R_1$) donne une cohérence à l'ensemble des résultats ($R_i$). Le problème du tiérisme est que ce 1/3 ne vient pas de n'importe où, il est évidemment issu de l'argument fréquentiste vu plus haut (un réveil sur trois est un réveil-face dans une longue répétition d'expériences) : les tiéristes doivent donc chercher un raisonnement où l'on aboutit à une probabilité épistémique 1/3, mais un

raisonnement guidé par l'intuition fondamentale d'une fréquence relative ou au moins d'une proportion (il y a trois compartiments dans le schéma de l'expérience mais un seul sur la ligne Face), et qui doit entrer en concurrence avec l'argument puissant de l'inertie doxastique, jusqu'à le vaincre, en tout cas montrer qu'il ne convient pas à une résolution de la Belle au bois dormant. Il est si difficile d'élaborer un tel raisonnement que l'on en rencontre aujourd'hui plusieurs, autrement dit coexistent plusieurs tiérismes, plusieurs tentatives pour montrer qu'au réveil la Belle doit accorder un crédit de 1/3 à face.

### 2.1. Le tiérisme de l'irrégularité bayésienne

Le tiérisme d'Elga et des chercheurs[130] qui, au début des années 2000, ont pris sa défense contre le demiste Lewis, n'est pas hostile au bayésianisme mais en souligne les limites actuelles. Une tragi-comédie en deux actes se joue. Dans le premier acte, la Belle entre dans l'expérience en estimant $C_{dim}(F) = 1/2$, et se réveille lundi en estimant $C_{rév}(F) = 1/3$ après avoir levé péniblement quelques obstacles. Dans le deuxième acte, quand on lui apprend qu'on est lundi, elle revient à $C_{lun}(F) = 1/2$. Cet acte du dénouement heureux est bayésien, la probabilité de FL et en conséquence celle de F augmentant simplement par conditionalisation sur FL $\vee$ PL.

C'est le premier acte qui fait mal au bayésianisme. La Belle est autorisée à changer sa croyance partielle parce qu'il lui arrive quelque chose que la théorie bayésienne ne sait pas traiter : elle perd une information ou, disons, elle est en manque d'information sur sa position temporelle, ce qui n'était pas le cas dimanche ; c'est lundi, mais elle l'ignore. La loi que l'on appelle principe d'inertie doxastique ne l'aide pas, à cause d'un point faible : elle présuppose qu'un savoir ne peut évoluer que par gain d'informations. Cependant, être autorisé à modifier une croyance,

[130] On peut citer Vaidman et Saunders (2001), Monton (2002) et Arntzenius (2002).

ce n'est pas encore être obligé. Or la Belle est passée d'une situation où une localisation temporelle est sans importance pour estimer la probabilité de face à une situation où au contraire la localisation temporelle est pertinente pour l'issue pile/face : dans l'expérience, en un jour impossible à dater, elle envisage des possibilités temporelles en plus grand nombre dans le cas où la pièce serait tombée sur pile. Elle doit s'adonner à un compte probabiliste neuf qui vraisemblablement ne conduira pas à la probabilité 1/2 du dimanche.

Pourquoi précisément 1/3 ? Une première ruse consiste à reconsidérer l'argument fréquentiste : en répétant l'expérience, des réveils-face-lundi auraient lieu en proportion 1/3 dans l'ensemble des réveils, de même pour les réveils-pile-lundi, de même pour les réveils-pile-mardi. Ainsi, la Belle a immédiatement une raison d'équilibrer les crédits qu'elle accorde à FL, PL et PM, alors qu'elle peine à trouver une raison pour briser cette équiprobabilité. Mais une autre ruse devrait confirmer qu'elle a vu juste : la Belle a la possibilité d'imaginer ce qui se passerait si on lui apprenait qu'on est lundi, deuxième acte qui selon elle se jouera ou pas aujourd'hui (qu'elle soit réveillée mardi est encore pour elle une possibilité) mais cela n'a pas d'importance. Si donc elle était informée qu'on est lundi, elle devrait, par habitude bayésienne, conditionaliser sur $FL \vee PL$ et par ce moyen nécessairement retrouver la probabilité de face 1/2 du dimanche. Un tiériste, quel qu'il soit, soutient en effet avec fermeté la démonstration faite plus haut du résultat ($R_2$) : $C_{lun}(F) = 1/2$. Et il n'est jamais mis en doute que $C_{lun}(F) = C_{lun}(FL)$. Or, en se servant entre autres de la définition de la probabilité conditionnelle, il est aisé de déduire de cette probabilité *a posteriori* la probabilité *a priori* $C_{rév}(F)$, même si cette méthode « à rebours » est inhabituelle :

$$\begin{aligned} 1/2 = C_{lun}(FL) &= C_{rév}(FL \mid FL \vee PL) \\ &= C_{rév}(FL \wedge (FL \vee PL)) \,/\, C_{rév}(FL \vee PL) \\ &= C_{rév}(FL) \,/\, (C_{rév}(FL) + C_{rév}(PL)). \end{aligned}$$

D'où $C_{rév}(FL) = C_{rév}(PL)$. Or $C_{rév}(PL) = C_{rév}(PM)$ et $C_{rév}(FL) + C_{rév}(PL) + C_{rév}(PM) = 1$. On en déduit $C_{rév}(FL) = 1/3$, puis via ($R_4$), qui est pour le tiérisme une évidence, on atteint $C_{rév}(F) = 1/3$, ce qu'il fallait démontrer[131].

### 2.2. Critique du tiérisme de l'irrégularité bayésienne

Dans le développement précédent, ce qui étonne et peut-être inspire peu confiance est l'asymétrie flagrante des deux « actes » interprétés par la Belle. Le scénario parait simple : l'agent perd une information temporelle, modifie sa croyance en face, premier acte ; il retrouve l'information, annule en conséquence sa modification doxastique, deuxième acte. Mais qu'il est difficile ce premier acte, qu'il est méfiant vis-à-vis du bayésianisme ! D'après lui, aucune règle de révision, aujourd'hui en tout cas, ne saurait transformer $C_{dim}(F) = 1/2$ en $C_{rév}(F) = 1/3$, cette dernière probabilité ne se découvre qu'en rusant, en faisant appel à l'argument fréquentiste, en *anticipant* une révision qui n'aura lieu que dans le deuxième acte. Un second acte simple comme bonjour, où soudain le bayésianisme a toutes les qualités requises, même lorsqu'il s'agit de manipuler des objets aussi nouveaux que les centres lewisiens !

En outre, si la Belle est véritablement un parangon de rationalité et si le tiérisme est bien l'idéal d'un tel parangon, elle ne peut pas échapper à une autre anticipation des événements et de ses croyances futurs, mais qui a lieu dimanche : elle sait ce jour-là, avant de s'engager dans l'expérience, qu'elle sera réveillée lundi et qu'elle accordera moins de crédit à face, en n'ayant pourtant reçu aucune information susceptible de modifier ce crédit, elle sait même qu'elle ne sera pas encore droguée, la première prise de la drogue à effet amnésique n'intervenant qu'après l'entretien avec les expérimentateurs. Pourtant, elle continue tout le dimanche à croire en face

[131] L'exposé de ce calcul n'est qu'une légère adaptation de celui d'Elga (2000).

au degré 1/2, non 1/3 : elle préfère donc obéir au principe principal de Lewis (un comble, puisque Lewis est demiste) plutôt qu'au principe de réflexion de van Fraassen. Elga est conscient de cela. Il retient cet énoncé du principe de réflexion :

> Un sujet, certain qu'il attribuera demain le degré de croyance *x* à la proposition R (à moins qu'il reçoive une nouvelle information ou subisse des incidents cognitifs entre-temps), doit *maintenant* attribuer le degré de croyance *x* à R.[132]

Elga pense que *Sleeping Beauty* fournit un contre-exemple à cette règle. Bien sûr, d'après David Lewis (2001), sa thèse tiériste ne met pas en défaut le principe de van Fraassen : c'est le principe qui met en défaut la thèse. Toutefois, nous pourrions répondre à ces deux philosophes que le conflit entre principe et thèse est discutable, car dès le lundi, la Belle a perdu la capacité de se repérer dans le temps au jour près. Sa mémoire n'a pas encore été altérée par la drogue, *mais elle n'en sait rien* ; il lui semble qu'elle est réveillée le lendemain du dimanche, mais elle se dit que c'est peut-être déjà mardi. Elle est en proie au doute et finalement dans le même état cognitif qu'après l'absorption de la drogue, comme si celle-ci avait un effet rétroactif. Les observateurs extérieurs savent qu'elle n'a pas subi un « incident cognitif », mais pour elle, subjectivement, être droguée ou non droguée est la même chose : elle est dans l'expérience, à l'intérieur d'une sorte de boîte, un lieu où les choses ne se passent pas comme à l'extérieur.

Inspiré par une publication du tiériste Frank Arntzenius, qui évoquait van Fraassen et les problèmes d'auto-localisation, Michael Titelbaum, qui est par ailleurs passionné par la Belle au bois dormant, a travaillé durant cette dernière décennie à un nouveau cadre bayésien prenant en compte les mouvements doxastiques et probabilistes dus à des *pertes* d'information, auto-localisante ou non. Il est notamment persuadé que de nouvelles règles telles qu'une conditionalisation généralisée et une réflexion généralisée

[132] Elga (2000), p. 146.

peuvent apporter des réponses aux paradoxes de l'auto-localisation[133]. Ces travaux novateurs n'ont pas encore marqué le débat autour de la Belle au bois dormant. Aussi, pour de nombreux tiéristes qui ne se contentent pas des résultats d'Elga, d'Arntzenius et d'autres critiques bienveillants de certains aspects du bayésianisme, la résolution du paradoxe passe par la recherche d'une information *reçue* par la Belle au moment de son réveil dans l'expérience, une information qui, d'une façon ou d'une autre, contraint l'agent à réviser sa croyance en face.

### 2.3. Le tiérisme fréquentiste ouvert

Le tiérisme du mathématicien Jean-Paul Delahaye a inspiré quelques philosophes français qui ont travaillé sur la Belle au bois dormant. Pour la première fois peut-être, Delahaye note l'importance pour la Belle de gagner et non perdre une certitude au réveil. La conjecture est hésitante mais sensible tout de même. La particularité intéressante de ce mathématicien est qu'il est un fréquentiste qui *s'ouvre* au raisonnement anthropique, à la théorie bayésienne et même au subjectivisme grâce à ses lectures de Leslie et de Bostrom notamment. Il évoque pourtant guère des degrés de croyance mais, un peu à la manière de Poincaré, ses « probabilités » sont plus ou moins objectives/subjectives. Il faut dire aussi qu'il écrit un article de vulgarisation scientifique, destiné à un public qui comprend souvent « fréquence » en entendant le mot « probabilité »[134].

Résoudre la Belle au bois dormant passe par une meilleure compréhension du phénomène d'anamorphose probabiliste, déformation de l'espace de probabilité par des effets de sélection observationnelle. L'effet

[133] Cf. Titelbaum (2008, 2013).

[134] Nous nous servons uniquement de Delahaye (2003) pour décrire ici sa manière de résoudre le paradoxe, mais nous assurons que le chercheur n'a pas produit de textes plus soignés et confidentiels.

le plus connu est l'effet de filtre ; il entre en jeu, par exemple, dans le problème compartimenté n° 3 du chapitre précédent, problème appelé « protocole sans mardi » par Delahaye, et ainsi schématisable :

| Jour / Tirage | **Lundi** |
|---|---|
| **Face** | 1 |
| **Pile** | 1/2 |

Engagée dans une longue série d'expériences de ce type, la Belle est un observateur en quelque sorte trompé sur le nombre de piles et de faces résultant des lancers d'une pièce équilibrée : en la laissant dormir une fois sur deux lors d'un tirage-pile, les expérimentateurs lui cachent, à elle seule, un certain nombre de ces tirages. Autrement dit, sa perception est soumise à un effet de filtre, certains tirages passent les mailles du filet de sa conscience. En même temps, la probabilité que la Belle associe à pile, à savoir 1/3, et qui n'est pas la probabilité estimée par les autres observateurs, n'est pas une erreur, elle est ancrée dans *son* vécu, *sa* réalité : si on lui demandait par exemple de parier sur pile ou sur face à chacun de ses réveils, elle s'enrichirait en pariant sur face.

Le bayésianisme n'est pas troublé par les effets de filtre. On comprend ici que la Belle qui se sait réveillée dans l'expérience « sans mardi » dispose d'une information qu'elle ne possédait pas avant l'expérience, et le principe de conditionalisation[135] permet de calculer efficacement des probabilités « déformées ». Le trouble arrive avec un autre genre d'anamorphose probabiliste, que notamment le protocole original de la Belle au bois dormant provoque chez l'agent.

[135] Peu familier du vocabulaire des théoriciens, Delahaye parle de la « formule de Bayes » et du « glissement bayésien ».

Revenons donc à la Belle au bois dormant originale. Pendant une longue série d'expériences répétées, la Belle est là encore comme isolée du monde extérieur, et elle regarde, elle seule, la probabilité objective 1/2 de pile à travers un miroir déformant. Cependant, cette distorsion de l'espace de probabilité est ici un *effet de loupe* : la Belle sait qu'elle existe en double, lundi *et aussi* mardi, dans les expériences-pile, aussi il lui apparait que la probabilité de pile grossit, comme une tête d'épingle vue à travers une loupe, et elle atteint 2/3 (elle doit être le double de la probabilité de face). Remarquons que la Belle n'est pas ici plus trompée ou moins clairvoyante que la Belle soumise à l'effet contraire, l'effet de filtre. Il n'y a pas, pour Delahaye, de raison de croire qu'elle est moins renseignée, plus ignorante que si elle participait à l'expérience sans mardi. Au réveil, la Belle apprend FL $\vee$ PL $\vee$ PM, une information temporelle dont l'obtention était prévue dès la veille de l'expérience, que l'on croirait incapable de modifier la probabilité de l'hypothèse éternelle F. Pourtant le mathématicien français la reformulerait ainsi : « Le protocole est maintenant enclenché, je suis dans une expérience où j'existe en double en cas de pile ». La certitude que l'on n'est pas en dehors mais dans l'expérience (dont on connait déjà les règles et l'effet grossissant) apparait indispensable au réajustement de la probabilité de face. On ne voit pas comment l'employer dans un calcul pour modifier des degrés de croyance, mais on ne peut pas nier qu'elle ait une pertinence ; simplement, on ne peut pas (ou on ne sait pas encore) conditionaliser sur elle.

Delahaye voit finalement l'anamorphose comme une généralisation de l'inférence bayésienne, elle traite des cas où conditionaliser semble impossible. Comprenons aussi que c'est grâce à un regard réaliste porté sur le dispositif *aléatoire*, non sur la seule pièce de monnaie mais surtout sur les réveils, et les réveils répétés comme dans les simulations fréquentistes, que la Belle doit estimer une probabilité de face de 1/3 au réveil, puis à

nouveau 1/2 si les expérimentateurs, en lui annonçant qu'on est lundi, la font sortir de sa boîte déformante.

### 2.4. Critique du tiérisme fréquentiste ouvert

En lisant Delahaye, qui ne distingue jamais deux facettes de la probabilité, nous pourrions croire qu'il les réconcilie efficacement. En réalité, malgré la profondeur indéniable de son propos, le mathématicien connait peu les raisonnements de ses adversaires dans la controverse *Sleeping Beauty*. Il affirme même, un peu tôt, que le paradoxe est résolu. Pour sûr, c'est Delahaye qui est résolu, résolu à se contenter de peu ; ce n'est évidemment pas un mal, mais ce n'est pas ce qu'on attend d'un réconciliateur du demisme bayésien et du tiérisme fréquentiste. Ainsi il convoque effets de sélection, probabilités changeantes selon les sujets, glissements bayésiens, mais le tiérisme fréquentiste est supposé à chaque pas de l'analyse : au fond il est tiériste parce que dans la répétition des expériences un réveil sur trois est un réveil-face, et même pas un réveil-face-lundi, car il n'est pas question d'envisager quelque chose comme un centre, ou de distinguer l'éternel et le temporel.

La méconnaissance du discours des théoriciens de l'auto-localisation peut jouer des tours à Delahaye. Il n'est pas étonnant qu'il sous-estime la résistance du paradoxe : il voit comme à l'ordinaire des mondes possibles, jamais des mondes centrés, et ce n'est pas parce qu'il aurait choisi de ramener des réveils séparés dans le temps à des réveils séparés dans l'espace logique, à la façon eccéitiste. Il semble même concevoir un réveil comme un événement parfaitement aléatoire, indépendant des autres réveils. Avec Léo Gerville-Réache, nous rappelons en 2015 dans l'article « Insaisissable Belle au bois dormant » que les réveils-pile-lundi et pile-mardi sont dépendants, nous montrons la difficulté de faire converger vers une probabilité une statistique basée sur l'observation d'une série de

réveils-face et de *couples* de réveils-pile, la probabilité-fréquence la moins contestable aux yeux de la théorie classique n'étant pas 1/3 mais 1/2, si l'on considère que les objets à compter ne sont pas des réveils solitaires. Delahaye envisage peut-être l'urne métaphysique dont nous avons déjà parlé, il voit peut-être un réveil de la Belle comme un réveil choisi aléatoirement parmi une multitude placés sur l'axe du temps, un tirage qui n'a pas à se demander *comment* ils se sont placés sur l'axe du temps. Mais il n'écrit jamais rien qui laisserait penser cela. Finalement, bien qu'il tente souvent de s'en protéger, il a tendance à confondre la Belle au bois dormant avec une variation où le jour du réveil serait décidé dans la nuit de dimanche à lundi par un second lancer de la pièce équilibrée qui complèterait le Tirage principal. Le tableau correspondant serait celui-ci :

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **Face(-face)** | 1 | 0 |
| **Face(-pile)** | 0 | |
| **Pile(-face)** | 1 | 0 |
| **Pile(-pile)** | 0 | 1 |

Mais il est possible de le réduire à cet équivalent :

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **Face** | 1/2 | 0 |
| **Pile** | 1 | |

Méconnu par les acteurs du débat, le mathématicien a un mérite : son effort réel pour pénétrer des théories qui lui étaient plutôt étrangères a porté

des fruits. Il apporte une réponse ferme à des demistes comme Leslie ou Franceschi (dont nous parlerons bientôt), tout en se permettant de ne pas être dépendant du tiérisme d'Elga. Avec ses maigres moyens, bien qu'il suggère que la bonne vieille conditionalisation montre ses limites face à la Belle au bois dormant, il prépare, par sa confiance en la pertinence d'une information temporelle acquise au réveil, un tiérisme qui soutient que le bayésianisme et le fréquentisme, contre toute attente, s'entendent sur $C_{rév}(F) = 1/3$.

### 2.5. Le tiérisme objectiviste

Un collectif d'une quinzaine de philosophes menés par John Pollock, parmi lesquels on trouve Paul Thorn et surtout Terry Horgan que nous allons bientôt revoir, a publié en 2008, sous le nom « The OSCAR Seminar »[136], un article qui prétend résoudre assez simplement la Belle au bois dormant en puisant dans la théorie objectiviste, non bayésienne. C'est un des derniers travaux de Pollock, décédé en 2009, et il est pour nous d'un grand intérêt : non seulement les auteurs formulent une information reçue par la Belle au réveil et susceptible de modifier son degré de croyance en face, mais encore ils proposent de l'intégrer à un raisonnement et un calcul rigoureux menant au degré 1/3. Cependant, il ne s'agit pas de modifier une probabilité *a priori* par conditionalisation sur l'information…

L'article distingue les deux versants de la probabilité. Il distingue plus exactement une probabilité indéfinie ou générale (notée p dans notre exposé), qui est, pour le dire vite, une « probabilité sur des propriétés de référence » qu'il invite à concevoir comme une probabilité-fréquence inadaptée au cas singulier, et une probabilité définie ou particulière (notée

[136] OSCAR est le nom d'un ancien projet de logiciel d'intelligence artificielle dirigé par Pollock.

C) qui peut être un degré de croyance. La seconde peut être dérivée de la première par une inférence directe qui prend cette forme :

(i) $p(Bx \mid Ax) = \rho$ (la probabilité (indéfinie) qu'un A soit B est ρ)
(ii) $A\lambda$ (λ est A)
(iii) $C(B\lambda) = \rho$ (la probabilité (définie) que λ soit B est ρ)

Soyons large et disons qu'un scénario de la Belle dure 72 heures, de dimanche midi à mercredi midi. Si $B(t, s)$ signifie « *t* est un instant dans un scénario *s* », si $T(x, s)$ signifie « *x* est le tirage à pile ou face du scénario *s* », si $Hx$ signifie « *x* amène face », alors on a :

$$p(Hx \mid B(t, s) \wedge T(x, s)) = 1/2$$

*x*, *t* et *s* sont des variables libres. Soient σ un scénario particulier et τ le tirage à pile ou face dans ce scénario σ. Dimanche soir, la Belle sait $B(\textit{maintenant}, \sigma) \wedge T(\tau, \sigma)$, *et rien d'autre qui puisse lui donner une probabilité différente*. À ce moment elle peut déduire par inférence directe que $C(H\tau) = 1/2$.

Imaginons que la Belle se réveille dans l'expérience, incapable de se souvenir d'un éventuel précédent réveil. En réfléchissant au fait qu'elle vient d'être réveillée, elle essaie d'estimer au mieux l'instant du réveil. Par exemple, elle estime que c'était il y a plus de neuf minutes mais moins de dix minutes. Elle suggère donc un intervalle de temps Δ durant lequel on l'a assurément réveillée. Si $W(t, s)$ signifie « La Belle a été réveillée dans le scénario *s* quelque part durant l'intervalle Δ (relatif à l'instant *t*) et ne s'est pas souvenu d'un précédent réveil dans *s* », si δ est la durée de Δ exprimée en heures, et si nous supposons une distribution uniforme des probabilités dans le temps, alors nous pouvons calculer :

$$p(W(t, s) \mid Hx \wedge B(t, s) \wedge T(x, s)) = \delta / 72,$$

$$p(W(t, s) \mid \neg Hx \wedge B(t, s) \wedge T(x, s)) = 2\delta / 72.$$

Remarquons que la seconde probabilité est le double de la première, puisqu'en cas de pile il y a deux intervalles de durée δ pour lesquels W(*t*, *s*) est vraie. Le théorème de Bayes et un calcul que nous ne détaillerons pas conduisent donc à un résultat prévisible :

p(H*x* | W(*t*, *s*) ∧ B(*t*, *s*) ∧ T(*x*, *s*)) = 1/3.

Voici que la Belle se réveille effectivement dans l'expérience. Elle sait B(*maintenant*, σ) ∧ T(τ, σ) et par conséquent pourrait directement inférer C(Hτ) = 1/2 comme le dimanche soir. Mais non ! La Belle apprend W(*maintenant*, σ), et W(*t*, *s*) ∧ B(*t*, *s*) ∧ T(*x*, *s*) est une propriété de référence logiquement plus forte, plus spécifique que B(*t*, *s*) ∧ T(*x*, *s*). C'est donc sur la base de notre dernier résultat que la Belle doit directement inférer C(Hτ) = 1/3. Il va sans dire qu'en C(Hτ) nous reconnaissons $C_{rév}$(F).

### 2.6. Critique du tiérisme objectiviste

Par le choix des variables, des propriétés, cette idée de l'intervalle Δ, cette utilisation de l'indexical de temps *maintenant*… la solution objectiviste est astucieuse. Peut-on être plus astucieux ? Joel Pust (2011) remarque que l'inférence directe au réveil, décrite par The OSCAR Seminar, est concurrencée par celle-ci :

(i) p(H*x* | T(*x*, *s*)) = 1/2
(ii) T(τ, σ)
(iii) C(Hτ) = 1/2

Nous aurions tendance à penser que W(*t*, *s*) ∧ B(*t*, *s*) ∧ T(*x*, *s*) est une propriété de référence logiquement plus forte que T(*x*, *s*). Seulement, un ensemble de triplets ne peut pas être un sous-ensemble d'un ensemble de paires, et l'ensemble des (*x*, *t*, *s*) qui ont la propriété de référence décrite par The OSCAR Seminar ne peut pas être un sous-ensemble de l'ensemble des (*x*, *s*) qui ont la propriété de Pust. Chercher ici ce qui est « logiquement

plus fort » n'a pas de sens dans la théorie objectiviste traditionnelle. Les deux inférences directes en conflit se mettent-elles mutuellement en échec ? Paul Thorn (2011) répond : non, à condition de repenser ce qu'il faut entendre par « logiquement plus fort ». Et le bon sens dicte de prime abord : si $n \geq m$, la $n$-propriété (propriété à $n$ emplacements) R est logiquement plus forte que la $m$-propriété Q si et seulement si est vérité logique que, $\forall\, x_1,\ldots, x_n$, $R(x_1,\ldots, x_n) \supset Q(x_1,\ldots, x_m)$.

Toucher à des définitions classiques comme le fait Thorn, c'est engendrer de nouvelles questions et complexifier une solution tiériste qui se voulait simple. C'est d'ailleurs la remarque que formule Kai Draper dans un tout récent article[137]. Ce philosophe pense aussi qu'on peut défier autrement l'argument objectiviste. Si L$t$ signifie « $t$ est un instant de lundi » et *cil* signifie « cet instant lundi », c'est-à-dire « lundi à l'heure (du jour) qu'il est maintenant », alors la Belle au réveil peut affirmer (i), (ii), et peut en déduire (iii) par inférence directe :

(i) $p(Hx \mid W(t, s) \wedge B(t, s) \wedge T(x, s) \wedge Lt) = 1/2$
(ii) $W(cil, \sigma) \wedge B(cil, \sigma) \wedge T(\tau, \sigma) \wedge Lcil$
(iii) $C(H\tau) = 1/2$

Non seulement elle peut mais elle doit, puisque selon les définitions les plus fiables de la « force logique », $W(t, s) \wedge B(t, s) \wedge T(x, s) \wedge Lt$ est plus forte (plus spécifique) que $W(t, s) \wedge B(t, s) \wedge T(x, s)$, que nous trouvons dans l'inférence de Pollock, laquelle serait ainsi mise en échec. Il faut quand même reconnaitre que l'objection de Draper est un jeu de logicien amoureux de paradoxes vicieux. Quand on la lit, une fois pour se mettre au courant, deux fois pour être sûr, trois fois pour être déconcerté, on a le même sentiment que face à l'argument de l'émeraude « vleue » de Goodman, à ceci près qu'on croit peut-être plus facilement Draper.

[137] Draper (2017).

Nous pensons que jouer avec le temps et les hypothèses temporelles peut engendrer des raisonnements faillibles, en tout cas contestables. Alors rivalisons nous aussi d'ingéniosité, mais sous un autre angle et en *déjouant* le temps, pour contester la solution objectiviste. La durée d'un scénario est sans importance pour The OSCAR Seminar : 50 ou 80 heures plutôt que 72 ne modifieront pas le résultat final. Certes. *Mais pourquoi un scénario devrait-il avoir la même durée en cas de pile et en cas de face ?* Changeons un peu les règles : la Belle participe à plusieurs expériences consécutives similaires à l'originale, mais sans jamais avoir de jour de repos et sans jamais rester endormie un jour entier. En effet, chacune des expériences prend fin au milieu de la nuit qui suit immédiatement le dernier réveil, et la suivante débute aussitôt, donc on relance la pièce et on réveille à nouveau, une ou deux fois, une Belle qui est toutefois toujours au courant du numéro de l'expérience (le nombre de lancers déjà effectués) grâce à un compteur installé dans sa chambre. Ainsi se succèdent des expériences de 24 ou de 48 heures. C'est une variante d'alembertienne dans le sens où un jour-face sur deux n'a plus d'existence, pour personne, pas seulement pour un sujet laissé dans l'inconscience ce jour-là : le temps se replie sur l'autre jour-face, premier et dernier de son scénario. Dans les conditions de cette variante, un tiériste guidé par la chaine des réveils-pile et des réveils-face croit toujours que la Belle qui se réveille au milieu d'une expérience doit estimer à 1/3 la probabilité de face si elle ne peut pas dater le jour. Pourtant, indéniablement, un scénario-pile dure un jour de plus qu'un scénario-face ; en raison du compteur de lancers, la Belle considère que se succèdent de tels scénarios, elle ne considère pas comme un scénario unique l'ensemble des expériences. Il apparait alors assez clairement que les calculs corrigés de Pollock mènent à une probabilité supérieure à 1/3, et peut-être à la probabilité demiste. L'arbitraire semble bien présent dans l'argument objectiviste. Un changement acceptable de quelques paramètres pourrait lui faire conclure des probabilités extravagantes.

### 2.7. Le tiérisme bayésien

Bien que Terry Horgan ait fait partie du collectif The OSCAR Seminar, il préfère et défend depuis plus de dix ans une autre solution tiériste du paradoxe qui a beaucoup fait parler d'elle. $C_{rév}(F) = 1/3$ serait le résultat d'une « mise à jour bayésienne synchronique »[138] d'une probabilité 1/2 *a priori*. Disons-le maintenant : Horgan ne fait pas allusion à des théories qui distinguent la mise à jour d'autres formes de réajustement des croyances, et on pourrait remplacer *updating* par un quasi-synonyme sans dommage. Plus tard, il fondra d'ailleurs en un seul mot les deux tiers de son expression et parlera de « conditionalisation synchronique »[139].

Horgan attire notre attention sur une variation du scénario de la Belle au bois dormant, imaginée par un autre philosophe[140], et très commentée dans la littérature. La Belle est réveillée le lundi *et le mardi quel que soit le résultat du lancer de la pièce*. Si pile, on lui administre à chaque fois la drogue à effet amnésique de l'expérience originale. Si face, on lui administre le lundi une drogue plus faible : en conséquence, la Belle est amnésique durant la première minute de son état de veille du mardi, puis elle recouvre la mémoire, et elle est alors sûre qu'on est mardi et que la pièce est tombée sur face. L'analyse de la variante nécessite de distinguer $C_{rév1}$, mesure de la croyance de la Belle durant la première minute de son état de veille, et $C_{rév2}$, mesure de sa croyance durant la deuxième minute. Il faut aussi distinguer non plus trois mais quatre compartiments associés aux hypothèses suivantes :

FL : « La pièce est tombée sur face et aujourd'hui est lundi »
FM : « La pièce est tombée sur face et aujourd'hui est mardi »

[138] L'expression n'est pas utilisée en 2004 dans le premier article de Horgan sur la Belle au bois dormant ; elle fait son apparition en 2007 dans un article plus mûr.

[139] Horgan et Mahtani (2013).

[140] Cian Dorr. Cf. Dorr (2002).

PL : « La pièce est tombée sur pile et aujourd'hui est lundi »

PM : « La pièce est tombée sur pile et aujourd'hui est mardi »

Dès l'instant où la Belle se réveille, ces quatre possibilités centrées sont pour elle équiprobables de toute évidence, donc $C_{rév1}(F) = C_{rév1}(FL) + C_{rév1}(FM) = 1/4 + 1/4 = 1/2$. Mais au bout d'une minute, si elle ne se souvient pas d'un précédent réveil, elle doit annuler le crédit qu'elle accordait à FM : $C_{rév2}(FM) = 0$. Selon Horgan et tous les tiéristes, rien ne vient favoriser ni discriminer une des trois hypothèses restantes, elles sont toujours équiprobables. La Belle en déduit $C_{rév2}(F) = C_{rév2}(FL) = 1/3$. Elle a diminué le degré de sa croyance en l'obtention de face, ce qui semble normal puisqu'elle l'aurait au contraire augmenté (jusqu'à 1) si elle s'était souvenu d'un précédent réveil.

La différence entre le problème original et cette variante est peut-être significative : dans l'original, si l'on excepte le moment où l'on apprend à la Belle qu'on est lundi, celle-ci n'a apparemment jamais l'occasion de conditionaliser sur une nouvelle information (telle que ¬FM dans la variante). Terry Horgan pense au contraire que la variante éloigne un peu dans le temps, ordonne et ainsi met en évidence certains mouvements dans l'esprit de la Belle, qui ont lieu simultanément dans l'expérience originale. Il comprend que la Belle ne reçoit aucune information éternelle entre dimanche et lundi, mais à un problème d'auto-localisation on doit habituer son esprit à répondre par des propositions auto-localisantes. Ainsi la clé de son analyse difficile est une proposition dont il souligne le caractère indexical, voire en certaines circonstances essentiellement indexical au sens de John Perry (qu'il préfère à David Lewis) :

V : « Je suis réveillée aujourd'hui par les expérimentateurs »

Le dimanche soir, le protocole de l'expérience originale en tête, la Belle peut déjà songer à ce qui va lui arriver, à ses futurs moments de veille et de sommeil, elle a conscience qu'elle peut être inconsciente mardi. Elle

sait qu'elle est réveillée ce dimanche mais pas par les expérimentateurs, elle sait que demain elle sera réveillée par ceux-ci, et ignore si après-demain elle le sera : V est au minimum hors de son domaine de certitudes, et même, n'existe pas *pour elle à ce moment précis*, si l'on considère que les mots qui la rattacheraient à V lui manquent à cause de son indexicalité. Bien sûr, elle estime à 1/2 la probabilité que la pièce de monnaie tombe sur face. Engagée dans l'expérience, voici que la Belle se réveille en perdant une information concernant sa localisation temporelle : elle ignore si ce jour est lundi ou mardi. Elle ignore également de quel côté est tombée la pièce. Aussi lui apparaissent quatre possibilités conformes à ses certitudes du dimanche, FL, FM, PL et PM, *essentiellement* indexicales, dans le sens où la Belle ne peut exprimer ce jour qu'à l'aide d'un indexical (« aujourd'hui »). Elle leur attribue la probabilité *a priori* $C_{rév}(FL) = C_{rév}(FM) = C_{rév}(PL) = C_{rév}(PM) = 1/4$. Elle est également sûre des quatre probabilités conditionnelles correspondantes : $C_{rév}(FM \mid V) = 0$, $C_{rév}(FL \mid V) = C_{rév}(PL \mid V) = C_{rév}(PM \mid V) = 1/3$. Or, la perte d'information est accompagnée par un gain : V prend tout son sens maintenant que la Belle est consciente à l'intérieur de l'expérience, V est une nouvelle certitude, une information elle aussi essentiellement indexicale, adaptée pour modifier les degrés de croyance précités. La Belle est en mesure de donner les probabilités *a posteriori* des quatre hypothèses, et elles sont, bien sûr, égales aux probabilités conditionnelles correspondantes. Il en résulte la nouvelle probabilité que la pièce soit tombée sur face : 1/3.

### 2.8. Critique du tiérisme bayésien

Horgan décrit là une révision doxastique qu'il assure être bayésienne mais est en tout cas très insolite, peut-être suspecte, parce qu'elle n'est pas diachronique comme l'est la conditionalisation habituelle : lors de l'émergence de la conscience de l'agent rationnel dans un environnement spécial qui déséquilibre les populations de ses parties temporelles dans les

mondes possibles, plusieurs événements mentaux se bousculent mais peu se succèdent dans le temps ; il y a des probabilités *a priori* et *a posteriori*, mais l'*a priori* et l'*a posteriori* ne sont plus l'avant et l'après et, si l'on peut dire, reprennent leur sens classique. La révision des probabilités des mondes centrés puis des mondes non centrés est synchronique, et pourtant ses étapes logiques font penser à la conditionalisation : *c'est* une autre conditionalisation dans l'esprit du tiériste.

Une perte d'information auto-localisante serait suivie logiquement et non chronologiquement par un gain d'information auto-localisante, dite synchronique : en perdant le jour où elle est actuellement consciente, la Belle a besoin d'envisager les compartiments associés à FL, FM, PL et PM ; en gagnant V, elle élimine la possibilité FM. Cet exposé des mouvements doxastiques donne l'impression de couper un cheveu en quatre, à tel point que leur synchronisme supposé ressemble à un aveu : on découpe logiquement et chronologiquement un mouvement doxastique simple qui est celui du demisme (nous y reviendrons), on s'arrange pour le recoller mais seulement chronologiquement pour en faire un mouvement doxastique tiériste. Certes, c'est une façon très sévère de formuler une objection à Horgan, elle sous-estime l'effort sincère que fournit ce chercheur pour faire comprendre ses vues.

Une « objection cartésienne » revient souvent sous la plume du double demiste Joel Pust[141]. Un agent rationnel évalue des probabilités épistémiques dans une situation épistémique *donnée* : lorsqu'il se réveille, il reçoit en premier, voire possède *a priori* cette certitude qu'il est réveillé parce que l'état de veille est condition de ses croyances et de ses connaissances ; qu'il exprime ou pas le jour par un indexical ne change rien. Pour la Belle engagée dans l'expérience, la probabilité de V est 1 d'emblée, ce qui lui interdit d'accorder le moindre crédit à l'hypothèse FM

[141] Pust (2008, 2013).

et donc ne lui donne aucune occasion de conditionaliser sur V, diachroniquement ou synchroniquement. En outre, quand il reste endormi, le sujet de l'expérience n'a évidemment jamais la possibilité de reconnaitre qu'il n'est pas réveillé, alors qu'il a toujours l'occasion, une fois conscient, d'exprimer le savoir selon lequel il est réveillé : l'expérience originale de la Belle est suffisamment différente d'une variante avec réveil mardi en cas de face pour que les conclusions de l'analyse de cette dernière ne soient pas transposables à l'originale[142]. On se demande comment la Belle pourrait conditionaliser sur V alors qu'elle n'a jamais la possibilité, ni de conditionaliser sur ¬V, ni de penser qu'elle pourrait bientôt conditionaliser sur V ou sur ¬V.

Horgan a plusieurs moyens de répondre : en considérant que les probabilités qu'il manipule présentent des similitudes avec les probabilités logiques[143] ; ou encore en expliquant que « aujourd'hui est lundi » a un pouvoir confirmant tel que $C_{rév}(F)$ ne peut qu'être inférieure à $C_{lun}(F) = 1/2$, cette dernière égalité étant d'ailleurs aussi défendue par Pust. Mais celui-ci, à ce propos, semble exprimer une crainte relative à l'utilisation par Horgan du vocabulaire de John Perry[144]. Comme beaucoup de théoriciens de l'auto-localisation, Perry s'était montré moins confiant que Lewis concernant le pouvoir réellement informatif ou la pertinence des « propositions à accessibilité limitée »[145], c'est-à-dire des propositions que seule une personne particulière peut saisir à un moment particulier. V et « aujourd'hui est lundi » sont de telles propositions. Horgan ne choisirait donc pas un bon mentor.

---

[142] Bradley (2003) présente pour la première fois cette objection, qui complète celle de Pust. Bradley l'adressait à Cian Dorr, non à Horgan.

[143] C'est l'ambiguïté que Pust (2008) relève dans la discussion avec son adversaire.

[144] Cf. Pust (2012).

[145] Perry (1993), p. 45.

## 2.9. La réparation tiériste d'une scission ontique-épistémique

Pust considère que Horgan propose « la meilleure défense de la position tiériste »[146] : le tiérisme bayésien serait plus évolué que les autres et donc plus attrayant. Nous avons des raisons d'être d'accord, même si peut-être, comme Pust, nous n'épousons pas le tiérisme.

Pour Horgan et Anna Mahtani, qui a adopté ses vues, il est impossible, si l'on est bayésien, de se dire que tout va bien dans la maison bayésienne mais que la Belle et quelques autres problèmes où interviennent des incidents cognitifs sont ingérables : ce serait dire que l'agent rationnel, prêt à être drogué ou déjà drogué, déraisonne, que son estimation de $C_{rév}(F)$ est le produit de sa déraison, et qu'on ne s'occupe pas d'un tel produit. Les deux tiéristes n'entendent donc pas les ruses d'Elga pour essayer de raccrocher les degrés de croyances de la Belle à des probabilités ontiques glanées pendant l'expérience ou à des croyances futures.

Toutefois, si son appartenance au collectif mené par Pollock est bien une preuve, Horgan n'est pas du tout hostile à la solution objectiviste, et c'est en cela qu'il réconcilie deux approches : la sienne, qui conclut une certaine probabilité épistémique à partir d'une autre probabilité épistémique, et une approche amie, qui fournit la même probabilité, mais entièrement fondée sur des probabilités ontiques. Que les raisonnements soient très différents importe peu. Ce philosophe est proche de la fin d'un mouvement tiériste à base fréquentiste qui cherche à s'allier le bayésianisme pour écarter toute dissonance ontique-épistémique. Nous percevons ce mouvement de deux façons. Par la progression principalement : quelque chose comme « le gain de l'information « Je suis réveillée aujourd'hui/maintenant » » est le résultat d'une prise de conscience de quelques années. Elga y est d'abord quasiment étranger, peut-être hostile. Delahaye, peu explicite, croit pourtant en un gain

[146] Pust (2013).

d'information qu'il formule très approximativement et ne sait pas utiliser dans un calcul. The OSCAR Seminar est plus précis dans la formulation et le calcul. Horgan est celui qui fait de l'information le cœur d'un nouveau processus pour mettre à jour une probabilité dont le caractère épistémique est enfin indéniable. Mais le mouvement tiériste est aussi synthétique : Horgan ne trouve que chez les défenseurs du tiérisme de l'irrégularité bayésienne la notion de perte d'information auto-localisante ; il ne trouve que chez les objectivistes l'idée d'une information indexicale synchronique. S'il connaissait plus directement le tiérisme fréquentiste ouvert au bayésianisme de Delahaye, il y verrait l'annonce d'une généralisation de la conditionalisation. Ce sont tous ces éléments « semi-bayésiens » organisés qui font la synthèse bayésienne du professeur de l'université d'Arizona.

La volonté de réparer une incroyable mésentente entre le fréquentisme et le bayésianisme comme s'il s'agissait d'aligner une probabilité épistémique sur une probabilité ontique, non pas dans une courte inférence formalisable mais dans un discours implicite liant des théories probabilistes rivales, est très sensible dans le courant tiériste. Explicité et résumé, ce discours tiériste serait : le fréquentisme nous donne p(F) = 1/3 (estimation de la Belle au réveil), un bayésianisme inquiétant nous donnait $C_{rév}(F)$ = l/2 ; le bayésianisme revisité « réaligne » sur 1/3 les deux genres de probabilité, soulagement ! Certaines solutions non tiéristes de la Belle au bois dormant donnent parfois aussi l'impression que les chercheurs veulent se rassurer face à l'étrange disharmonie ontique-épistémique.

## 3. Le demisme

Retournons aux quatre relations ($R_i$). Pour rendre leur ensemble cohérent, on peut, contrairement aux tiéristes, accepter et défendre

($R_1$) $C_{rév}(F) = C_{dim}(F)$, c'est-à-dire $C_{rév}(F) = 1/2$, mais en revanche contester le seul résultat ($R_2$) $C_{lun}(F) = 1/2$, lui préférer $C_{lun}(F) = 2/3$ et ainsi prendre parti pour le demisme, aussi appelé simple demisme parce qu'il répond 1/2 seulement à la question principale des expérimentateurs (alors qu'une position sœur, le double demisme, répond 1/2 aux deux questions). Le principal problème du demisme est celui de la justification de $C_{lun}(F) = 2/3$, probabilité fortement contrintuitive qui dissuade beaucoup de chercheurs de suivre le chemin montré par feu David Lewis. Nous distinguons deux simples demismes : un demisme bayésien véritablement lewisien et un demisme ouvert au fréquentisme.

### 3.1. Le demisme bayésien

Premier en date, le demisme bayésien ne s'occupe tout simplement pas de critiquer la simulation mentale consistant à répéter l'expérience et enchainer les réveils. Les fréquences ne concernent pas les rares avocats de ce demisme ; les propensions, oui, ponctuellement. Tout ou presque est une affaire de croyances dynamiques.

Reformulons l'argument de D. Lewis (2001). C'est le dimanche soir. Avant de s'endormir, la Belle comprend l'indexicalité de (FL $\vee$ PL $\vee$ PM) : elle ne « recevra cette information » qu'à un certain moment, dans l'expérience, mais avant l'expérience elle ne peut que *s'attendre* à la recevoir. Voici qu'elle se réveille dans l'expérience. Elle reçoit (FL $\vee$ PL $\vee$ PM) et n'en est pas surprise, donc cette information n'a aucune force. La Belle n'apprend absolument rien de pertinent pour l'issue pile/face. Ajoutons qu'à cause de la drogue qui efface une partie de ses souvenirs, elle ne peut pas distinguer un simple réveil d'un double réveil, donc être réveillée deux fois et informée deux fois en cas de pile est pour elle sans importance. En vertu du principe d'inertie doxastique, $C_{rév}(F) = C_{dim}(F) = 1/2$. Elle déduit de ce résultat que $C_{rév}(FL) = 1/2$, $C_{rév}(PL) =$

$C_{rév}(PM) = (1 – 1/2) / 2 = 1/4$. La Belle est maintenant dans l'attente d'une autre information, mais elle ignore si celle-ci viendra : les expérimentateurs pourraient en effet lui annoncer qu'on est lundi, ou pas (dans ce dernier cas, avant de se rendormir, elle saurait qu'on est mardi et que la pièce est tombée sur pile). Voici que les expérimentateurs font l'annonce : la Belle est cette fois-ci surprise par $FL \vee PL$[147]. Une simple conditionalisation sur cette information permet de trouver :

$$\begin{aligned} C_{lun}(FL) &= C_{rév}(FL \mid FL \vee PL) \\ &= C_{rév}(FL).C_{rév}(FL \vee PL \mid FL) / C_{rév}(FL \vee PL) \\ &= 1/2 \,.\, 1 / (1/2 + 1/4) = 2/3. \end{aligned}$$

Enfin, $C_{lun}(F) = C_{lun}(FL) = 2/3$ en raison d'une équivalence logique des deux hypothèses. La Belle est en mesure de répondre à la question subsidiaire de ses gardiens. Elle perçoit face comme plus probable que pile ; c'est contrintuitif, mais l'intuition nous leurre parfois.

Pourtant, l'argument qui soutient le résultat $(R_2)$ $C_{lun}(F) = 1/2$ parait inattaquable. En apprenant qu'on est lundi, la Belle n'est plus différente d'un observateur extérieur qui aligne $C_{lun}(F)$ sur la propension de la pièce à tomber sur face. Remarquons que le déroulement de l'expérience ne dépend pas du Tirage tant que mardi matin n'est pas venu. Imaginons un scénario similaire où la pièce n'est lancée que dans la nuit de lundi à mardi afin de déterminer si un autre réveil doit avoir lieu avant que l'expérience ne prenne fin : cette légère variation des règles, connue par la Belle, ne peut pas changer ses réponses et en particulier, quand elle est réveillée lundi et qu'elle en est informée, elle doit croire que la pièce qui sera lancée dans quelques heures a une chance sur deux de tomber sur face. On peut même l'inviter à lancer elle-même la pièce, puisque de toute façon la drogue va

[147] Darren Bradley dirait que la Belle *découvre* $FL \vee PL$, c'est-à-dire qu'elle prend connaissance de la vérité d'un contenu dont la valeur de vérité n'a pas changé dans la période qui environne la découverte.

lui faire oublier la journée du lundi : une fois que la pièce est dans sa main, la Belle peut difficilement estimer à 2/3 la probabilité d'obtenir face. Ce serait violer le principe principal.

Le prolifique demiste bayésien Darren Bradley a récemment développé une argumentation qui, selon lui, clarifie la critique trop hâtive que Lewis opposait à cet argument ennemi du 2/3[148]. Le principe principal est complexe. Quand on annonce à la Belle qu'on est lundi, celle-ci acquiert en fait une information inadmissible[149], qui peut justifier que ses degrés de croyance s'écartent des chances objectives. Un exemple plus clair est celui d'une boule de cristal infaillible qui prédit qu'une pièce donnée va tomber sur pile. Tout agent ayant cette prédiction en tête est certain que la pièce va tomber sur pile : une telle information sur le futur disqualifie les probabilités habituellement associées à l'objet « pièce de monnaie ». Autre exemple extrême : la Belle qui apprend qu'on est mardi devient tout simplement certaine que la pièce est tombée sur pile, donc ne fait plus coïncider probabilités épistémique et ontique. Puisque « on est mardi » est une information inadmissible, alors l'information concurrente « on est lundi » l'est aussi et justifie que la Belle puisse se désolidariser de la probabilité 1/2 en adoptant à la place 2/3. L'annonce « on est lundi » ressemble à une prédiction : jusqu'alors incapable de se repérer dans le temps, un agent qui soudain apprend la date ou l'heure apprend en même temps, en quelque sorte, des informations sur le futur, puisqu'il découvre quelles sont, parmi toutes les positions temporelles possibles, ou plutôt parmi toutes les parties temporelles de l'agent, celle qui est actuelle, celles qui appartiennent au passé et celles qui appartiennent à l'avenir.

---

[148] Avec Léo Gerville-Réache, nous avons résumé en 2015 dans notre « Insaisissable Belle au bois dormant » le long développement de Bradley (2011b). Ce résumé est ici repris avec des modifications légères.

[149] Inadmissible au sens donné par Lewis dans son *Subjectivist's Guide*. Cf. le chapitre 2 de notre thèse.

### 3.2. Critique du demisme bayésien

Nous constatons d'abord, en le comparant à la mise à jour synchronique du tiérisme bayésien, la simplicité du mouvement doxastique de la Belle au réveil selon Lewis : seules des croyances temporelles sont révisées, comme si le sujet n'avait fait qu'avancer dans le temps, sans rencontrer d'influences affectant la probabilité 1/2 de face, parce que les événements de l'expérience étaient attendus dimanche. Pourtant, il y a des situations où il n'est pas du tout clair que la réalisation d'un événement prévu laisse inchangées des croyances éternelles. Un agent rationnel va observer trois flashs de lumière rouge ou verte, l'un après l'autre. Il sait qu'auparavant ses amis auront lancé une pièce équilibrée et produiront un flash rouge et deux verts en cas de face, deux flashs rouges et un vert en cas de pile ; il sait aussi que chaque flash a pour effet spécial de se faire oublier de l'observateur quelques secondes après, et avant la venue éventuelle d'un autre flash[150]. Voici que l'agent, qui ne se souvient d'aucun précédent flash, voit un flash rouge. Avant d'oublier ce flash, à quel degré doit-il croire que la pièce est tombée sur face ? Peut-être bien 1/3. Pourtant ce flash rouge était prévu, il n'est pas vraiment ce que nous appellerions une surprise. Un demiste dira peut-être que dans ce protocole l'agent qui attend un flash ne s'attend pas à voir nécessairement un rouge, sa mémoire ayant déjà pu être altérée : il y aurait donc surprise. Mais à ce moment-là, la Belle qui se réveille et comprend que l'expérience est en cours est elle aussi surprise de constater qu'elle est dans l'expérience (lundi ou mardi), c'est-à-dire qu'elle n'en est pas sortie, que ce jour n'est pas mercredi. En effet, mercredi elle a oublié son ou ses précédents réveils, elle est amnésique comme si l'expérience était encore en cours, et c'est par d'autres moyens qu'elle comprend que son aventure est bel et bien terminée.

[150] Nous empruntons ce scénario des trois flashs à Weintraub (2004).

Supposons qu'un demiste considère toujours (FL ∨ PL ∨ PM) comme une non-surprise même après cette objection ; il faudrait alors qu'il réexplique plus profondément pourquoi il identifie à une surprise l'annonce « on est lundi » et la nouvelle question des expérimentateurs, ces événements prévisibles arrivant à chaque expérience. Il expliquerait peut-être que c'est une non-surprise dans le contexte d'une expérience menée sur plusieurs jours, mais une surprise dans le contexte réduit du réveil isolé présent. Nous continuerions alors à objecter : la Belle ne reçoit donc pas (FL ∨ PL ∨ PM) dans un réveil isolé ? On a l'impression que certains tiéristes ont finalement une bonne approche de la notion de « surprise » lorsqu'ils reconnaissent que la Belle au réveil, certes, savait déjà qu'elle serait réveillée pendant l'expérience, puis qu'ils expliquent plus subtilement qu'elle ne savait pas qu'elle serait réveillée *maintenant*. Ici l'indexical essentiel ne ferait peut-être pas *réagir* l'agent, comme chez Perry, mais agiterait ses croyances éternelles via ses croyances temporelles (F via FL).

Venons-en à la défense de $C_{lun}(F) = 2/3$ par Darren Bradley, qui fait de notre pouvoir de localisation dans le temps une prescience insoupçonnée discréditant des propensions objectives. Son argument est très contestable. Le protocole de l'expérience de la Belle interdit un réveil-face-mardi, autrement dit la propension de ce dispositif aléatoire à faire advenir un réveil-face le mardi est nulle, ou encore : la Belle a toutes les chances *objectives* d'être dans un monde-pile quand elle est réveillée mardi. Apprendre qu'on est mardi la rend certaine que pile est venu, mais pourquoi ? Parce qu'elle calque ses degrés de croyance sur d'autres chances objectives que celles associées à une simple pièce de monnaie. S'il en est ainsi le mardi, il en est de même le lundi, et la propension du dispositif expérimental à faire advenir un réveil-face le lundi (qui se confond dans ce cas précis avec la propension de la pièce à tomber sur face) est 1/2 et non 2/3. Bradley se focalise sur la pièce comme si elle

représentait tout le dispositif aléatoire : c'est un angle de vue trop étroit. Il oublie que les événements produits sont des réveils et ne considère comme « chances » que la propension de la pièce à tomber sur un côté plutôt que sur l'autre. De plus, il croit que les informations inadmissibles se rencontrent facilement dans la vie d'un agent rationnel. En réalité, le demiste relève des écarts entre probabilités épistémiques et ontiques dont il a lui-même forcé l'existence, soit en occultant des éléments du dispositif aléatoire, soit en brodant une histoire pénétrée par le surnaturel, comme celle de la boule de cristal.

### 3.3. Le demisme ouvert au fréquentisme

Pour certains demistes, les tiéristes sont des fréquentistes imprudents. Le fréquentisme n'est ni une mauvaise approche, ni le point faible de la solution bayésienne demiste, il doit au contraire démontrer la supériorité de cette solution. Un théoricien objectiviste doit pouvoir adhérer au demisme comme, déjà, un subjectiviste le peut. En réaction au comptage fréquentiste des réveils élémentaires, ces demistes ouverts considèrent des séries de réveils, ils groupent les réveils dans l'unité de l'expérience. Cela leur permet d'analyser la Belle au bois dormant sans les difficultés liées à l'auto-localisation dans le temps.

John Leslie est d'avis que le tiérisme qu'il combat nait à la pensée de la répétition de l'expérience. Cette simulation mentale est salutaire dans la plupart des cas, pas avec la Belle au bois dormant. En effet, lorsque l'expérience est reproduite de très nombreuses fois, les réveils-pile comme les réveils-face deviennent *effectifs* et on trouve une majorité de réveils-pile : les résultats tiéristes semblent alors l'emporter. Mais si l'on est sûr qu'elle n'a lieu qu'une fois, alors un réveil-face est seulement *possible*, et ce n'est pas un réveil-pile mais *une paire* de réveils-pile qui est, elle aussi, seulement *possible*. Dans ce cas, la Belle qui, d'une part, sait que la pièce

de monnaie décide équitablement s'il y aura un seul réveil ou une série de deux réveils, et qui, d'autre part, n'a pas le souvenir d'avoir déjà été réveillée au cours de l'expérience, n'a pas de raison de croire qu'être dans le monde-pile est plus probable qu'être dans le monde-face. Et le fait qu'en cas de pile il y ait deux entretiens et donc que la question lui soit posée deux fois (on ne compte pas la question subsidiaire du lundi) n'est pas perturbant[151].

Avant d'évoluer vers la solution de la désambiguïsation que nous détaillerons bientôt, Paul Franceschi était un demiste influencé par Leslie mais pas convaincu par sa distinction expérience-unique - expériences-renouvelées. La répétition ne doit pas être un obstacle : il faut remarquer la dépendance des réveils-pile et considérer des classes de référence de réveils groupés. Le philosophe français résume ainsi sa manière de corriger ce qui est pour lui l'erreur tiériste :

> On ne peut pas additionner les réveils-face le lundi et les réveils-pile le lundi, car il ne s'agit pas du même objet. Les réveils-pile le lundi sont indissociables des réveils-pile le mardi : on ne peut avoir un réveil-pile le lundi sans un réveil-pile le mardi. Pour cette raison, alors que les réveils-face le lundi comptent 1 (objet), les réveils-pile le lundi et les réveils-pile le mardi ne comptent qu'1/2 (1/2 objet).[152]

Un réveil-face n'est pas le même objet qu'un réveil-pile car c'est un groupe dont la particularité est de n'avoir qu'un membre ; c'est une série au même titre que la série des deux réveils-pile indissociables. Franceschi ne critique pas le principe du comptage fréquentiste, il reprécise ce que doivent être les objets comptés : ses groupements d'un ou de deux réveils ne sont pas arbitraires mais au contraire exigés par le protocole. De son

[151] Jean-Paul Delahaye nous a fait parvenir sa correspondance avec Leslie, où celui-ci livre son analyse, laquelle n'a jamais été publiée. Les échanges de courriels entre les deux chercheurs datent de 2003.

[152] Extrait d'une correspondance personnelle.

point de vue, même si l'expérience était répétée, la Belle incapable de se repérer dans le temps devrait toujours estimer à 1/2 la probabilité de face. Seul le gain de l'information « on est lundi » peut faire évoluer cette probabilité.

Le philosophe américain Roger White sait que les tiéristes sont préoccupés par le difficile problème de l'information nouvelle acquise au réveil. Or le demisme, lui, n'a aucun mal à trouver une information efficace dans des variantes compartimentées :

W : « Je suis réveillée au moins une fois durant l'expérience »

Considérons par exemple :

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **Face** | 1/2 | 0 |
| **Pile** | 1/2 | 1/2 |

Ici la Belle peut ne pas être réveillée durant l'expérience. Lorsqu'il y a réveil, W, qui est éternelle, semble plus appropriée que toute autre information pour modifier des croyances éternelles, par la même révision bayésienne que celle qui est opérée dans des milliers de problèmes sans auto-localisation. Si la Belle conditionalise sur W au réveil, elle peut calculer $C_{rév}(F) = 2/5$. Dans le cas du problème original où la Belle est toujours réveillée, W laisse évidemment la probabilité de face du dimanche inchangée[153]. Il est inutile d'expliquer davantage le défi que White lance aux tiéristes, il suffit de constater qu'à une localisation du sujet dans un jour inconnu, il oppose une localisation dans l'expérience connue : « au moins une fois durant l'expérience » est la manière circonspecte de dire

[153] Nous ne gardons ici qu'une partie de l'analyse de White (2006).

simplement « dans l'expérience ». Voilà pourquoi White, bayésien de prime abord, peut tout de même être rangé parmi les demistes ouverts : il sent avec eux l'incassable unité de l'expérience et délaisse l'indexicalité des « aujourd'hui » et des « maintenant » dont les tiéristes aussi bien que les demistes fermés au fréquentisme ne peuvent se passer.

### 3.4. Critique du demisme ouvert au fréquentisme

Les demistes ouverts posent une question très importante dont les doubles demistes et les désambiguïseurs se souviendront : en cherchant à se repérer dans le temps, un sujet se situe-t-il plus facilement dans un jour qu'il ne peut pas dater que dans l'expérience datable ? Ils montrent qu'établir une probabilité-fréquence est moins simple qu'il n'y paraissait lorsque le caractère élémentaire d'un réveil en faisait l'objet à compter par excellence. Si une localisation dans le jour correspond à un comptage de réveils élémentaires menant aux résultats tiéristes, une localisation dans l'expérience correspond à un comptage de séries de réveils consolidant le demisme.

L'immense et fâcheux défaut du demisme ouvert, c'est qu'il laisse au demisme bayésien le soin de démontrer que $C_{lun}(F) = 2/3$. Cette probabilité contrintuitive est pourtant le grand problème, nous le savons. Cela dit, reconnaissons que Franceschi a produit un texte sur le sujet[154], où il défend notamment une analogie entre la Belle au bois dormant et un problème avec une urne un peu spéciale. L'urne contient une boule rouge et une boule verte, mais cette dernière a la propriété d'être insaisissable si et seulement si la pièce équilibrée lancée secrètement avant le tirage d'une boule est tombée sur face. En plongeant sa main dans l'urne et en saisissant

[154] Il s'agit d'un texte de 2005 intitulé « *Sleeping Beauty and the Problem of World Reduction* ». C'est une ancienne version de Franceschi (2010), aux différences significatives.

une boule, un agent rationnel doit croire au degré 1/2 que la pièce est tombée sur face, au degré 3/4 qu'il va tirer une boule rouge, au degré 1/4 qu'il va tirer une boule verte, des probabilités conformes aux fréquences observées en répétant virtuellement le tirage de très nombreuses fois. Ces chiffres rappellent les résultats demistes $C_{rév}(F) = 1/2$, $C_{rév}(FL \vee PL) = 3/4$ et $C_{rév}(PM) = 1/4$, et la boule verte insaisissable une fois sur deux est censée représenter le sommeil prolongé de la Belle le mardi en cas de face. Nous comprenons que tirer une boule puis constater qu'elle est rouge est pour Franceschi analogue à se réveiller puis apprendre que c'est lundi. Or, après avoir tiré une boule et constaté qu'elle est rouge, l'agent augmente jusqu'à 2/3 le crédit qu'il accorde à face, soit en conditionalisant, soit en faisant une nouvelle fois confiance à des probabilités ontiques objectives. On retrouverait donc $C_{lun}(F) = 2/3$. L'analogie échoue pourtant, Franceschi l'a d'ailleurs abandonnée[155] en comprenant le souci : ce problème d'urne est en réalité analogue au problème compartimenté ainsi schématisable :

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **Face** | 1 | 0 |
| **Pile** | 1 | |

Le demiste français a commis une erreur similaire à celle du tiériste Jean-Paul Delahaye[156] : à force de se cacher l'auto-localisation, il oublie un court instant que plusieurs réveils peuvent avoir lieu au cours d'une

[155] Plus exactement, en évoluant vers une résolution désambiguïsatrice du paradoxe, Franceschi a modifié le problème d'urne pour défendre une analogie qui sert ses vues nouvelles.

[156] Rappelons que, chez Delahaye, la première case du tableau qui a tendance à parasiter le tableau résumant le problème original contient la probabilité 1/2, et non 1 comme ici.

expérience unique. Montrer ou rendre plausible $C_{lun}(F) = 2/3$ à l'aide d'un dispositif aléatoire tel qu'une urne est peut-être une mission impossible.

Il est important de dire un mot du défi de White. Un vrai défi lancé aux tiéristes. Il nous rappelle évidemment l'analyse de Terry Horgan et promet les étincelles d'un combat entre :

V : « Je suis réveillée aujourd'hui par les expérimentateurs »

W : « Je suis réveillée au moins une fois durant l'expérience »

Le conflit féroce des intuitions tiériste et demiste apparait peut-être plus nettement dans une variation de ce genre :

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **Face** | 1 | 0 |
| **Pile** | 2/3 | 2/3 |

Ici, un réveil de la Belle au cours de l'expérience devrait confirmer pile selon un tiériste qui constate que la somme des probabilités sur la ligne Pile est supérieure à la somme des probabilités sur la ligne Face ; mais il devrait confirmer face aux yeux d'un demiste qui constate que la Belle n'est certaine d'être réveillée dans l'expérience qu'en cas de face ! Horgan répond à White que son information V est gagnante, mais il rencontre des difficultés[157]. Dans le cas de la Belle au bois dormant originale, W ne modifie pas la croyance en face du dimanche, seule V la modifie. Pourtant, on imagine aisément des variantes où seule W a la force d'agiter des croyances. L'avantage de V est qu'elle est plus spécifique dans le sens où elle semble impliquer W ; nous détenons apparemment une information plus fiable quand nous réussissons à nous localiser dans un jour précis

[157] Horgan (2007).

plutôt que dans un groupe de plusieurs jours. Le souci est que V ne localise pas la Belle lundi ou bien mardi, mais *aujourd'hui*. Il faut donc encore être conquis par l'optimisme de Horgan vis-à-vis des indexicaux essentiels pour préférer V à W.

### 3.5. La réparation demiste d'une scission ontique-épistémique

Des demistes comme Lewis et Bradley se désintéressent du fréquentisme mais pas du propensionnisme. Néanmoins c'est très épisodiquement qu'ils font appel aux propensions, et ils ne justifient que rarement des alignements entre probabilités ontiques et épistémiques. Ils assument même des écarts entre les deux genres, tant ils ont foi dans le bayésianisme. Ils ne partagent donc pas l'inquiétude d'Elga devant la divergence étonnante du tiérisme fréquentiste primitif et du demisme bayésien primitif. C'est avec Leslie que cette inquiétude s'empare du courant demiste. Le spécialiste du *Doomsday Argument* ne se laisse pas abattre, il porte un regard bienveillant sur le fréquentisme tout en relevant un biais qui était invisible à nos yeux avant l'avènement des problèmes d'auto-localisation et d'auto-sélection (du raisonnement anthropique). Lorsque l'expérience de la Belle est répétée, les fréquences relatives de réveils guident assurément vers de justes probabilités… mais qui ne sont valables que dans cette situation de répétition. C'est d'ailleurs aussi l'avis du double demiste Nick Bostrom, autre spécialiste du raisonnement anthropique. Lundi, mardi ou aujourd'hui sont des moments du monde et non des états du monde, et ça change tout. Lorsque plusieurs réveils-pile et réveils-face sont effectifs en ce monde et étalés dans le temps, l'agent réveillé peut donner un sens à une probabilité de vivre actuellement tel moment-face. Mais l'expérience unique, non répétée de la Belle ou une de ses variantes compartimentées proches se distingue parce qu'elle n'offre l'effectivité qu'à un certain nombre de réveils-pile *ou* (exclusif) à un certain nombre, éventuellement différent, de réveils-face : les probabilités

naturelles sont alors les probabilités de mondes possibles (non centrés). Ainsi, d'après Leslie, le fréquentisme est grand quand il comprend que le passage du cas répété au cas unique, sans conséquence sur les probabilités dans la plupart des problèmes, ne doit pas pour autant devenir un réflexe gommant des subtilités.

Toutefois, en se privant de fréquences relatives pour résoudre précisément la Belle au bois dormant (unique), Leslie ne propose pas de solution qui alignerait des probabilités épistémiques sur des probabilités ontiques conformes aux réponses demistes. Seul Franceschi tente de le faire, mais il échoue devant une difficulté de taille : justifier ontologiquement la probabilité 2/3 de face sachant lundi, « prédite » par les demistes bayésiens. White, quant à lui, est celui qui exprime le plus clairement l'unité de l'expérience datable, plus forte que l'expérience décomposable en réveils non datables. Il couronne la prise de conscience progressive de la possibilité d'un argument demiste qui regarde bien en face un problème d'auto-localisation subtil et déroutant et le résout en extrayant de lui le problème ordinaire qu'il pourrait être et sur lequel fréquentistes et bayésiens pourraient s'entendre.

La *volonté* de réparer la scission ontique-épistémique décelée dès le premier article d'Elga en 2000 est repérable chez les demistes ouverts au fréquentisme. La réparation en elle-même n'est pas manifeste chez eux. On sent qu'il manque une dernière étape synthétisant les qualités des raisonnements des divers auteurs. Le tiérisme trouve en l'analyse de Horgan une étape décisive, alors que le demisme attend et attendra peut-être longtemps, car dans les quelques publications demistes, le bayésianisme domine largement. Il est à craindre que $C_{lun}(F) = 2/3$, résultat alternatif à $(R_2)$, soit dans l'avenir, tant qu'il n'aura pas d'assise ontologique, éminemment suspect aux yeux des non-demistes.

## 4. Le double demisme

Tout aussi soucieux de cohérence, certains chercheurs acceptent les relations ($R_i$) sauf ($R_3$) $C_{lun}(FL) > C_{rév}(FL)$, qu'ils remplacent par $C_{lun}(FL) = C_{rév}(FL)$. Ils estiment que la Belle doit répondre 1/2 à la question principale de l'entretien, puis encore 1/2 à la question subsidiaire du lundi, d'où leur nom courant de doubles demistes. Prises isolément, ces probabilités sont pour eux les plus intuitives et résultent des meilleures parties des raisonnements demistes et tiéristes. Lewis a raison de suivre un principe d'inertie doxastique au réveil ; Elga a raison de fonder le degré de croyance de lundi sur des chances objectives. Seulement, on ne soude pas les deux raisonnements sans engendrer au pire un monstre, au mieux quelque chose de très suspect. Ici, lorsque l'hypothèse PM est disqualifiée par l'annonce « on est lundi », sa probabilité n'est pas redistribuée sur FL et sur PL, mais seulement sur PL. Le principe de conditionalisation n'est pas utilisé, est peut-être violé. On a beau se dire que tout est possible avec l'auto-localisation, on est quand même en présence d'une bizarrerie d'autant plus folle qu'aucun double demiste n'est hostile au bayésianisme.

### 4.1. Le double demisme du bayésianisme adapté

Nous nous remémorons le chapitre 2, l'imaging, l'analyse du problème de la machine à dupliquer… Nous nous doutons bien que c'est de cela qu'il s'agit. En effet, mais ça ne concerne pas tous les doubles demistes, seulement ceux qui pensent que la Belle suit une règle de révision doxastique nouvelle adaptée aux croyances auto-localisantes, révise les crédits qu'elle attribue à *certains* possibles mais, ce faisant, conserve la probabilité épistémique de FL et donc de F.

Chris Meacham est le premier à intervenir[158]. Ce grand lecteur de Lewis ne semble pourtant pas avoir l'imaging en tête. Il revient sur la célèbre phrase de Lewis qui se demandait ce qu'il faut modifier dans la théorie de la décision quand on veut prendre en compte toutes les croyances *de se* et pas seulement les *de dicto* :

> Réponse : très peu. Nous remplaçons l'espace des mondes par l'espace des mondes centrés, ou par l'espace de tous les habitants des mondes. Tout le reste est comme avant.[159]

Cette réponse est pour Meacham inconsidérée. Tout le reste ? Sûrement pas. Nos certitudes sont permanentes quand nous révisons nos croyances en suivant le principe standard de conditionalisation : une proposition que nous tenons pour vraie avant la révision, nous la tenons pour vraie après. Or, nous ne sommes pas toujours certains qu'il est midi (que nous sommes au temps midi), mais nous le sommes un court instant quand nous regardons une montre qui donne l'heure exacte et que nous constatons qu'il est midi. Une minute plus tard, nous sommes certains qu'il n'est pas midi. Cette simple remarque laisse penser que la dynamique des croyances *de se* est sophistiquée.

Le temps qui passe emporte naturellement les croyances de qui se localise dans le temps, mais pas les croyances éternelles, c'est-à-dire les croyances de qui se localise dans des mondes possibles. Il arrive des situations où nous envisageons des populations différentes (d'un monde possible à un autre) de nos parties temporelles, mais là encore il est naturel de penser que nos croyances qui nous localisent dans des mondes ne souffrent pas d'un tel déséquilibre. Lorsque la Belle passe de dimanche à lundi, elle passe aussi d'une situation d'équilibre des populations de possibles à une situation de déséquilibre, et sa croyance en face n'est pas

[158] Une première version « *preprint* » de Meacham (2008) circulait depuis 2003.

[159] D. Lewis (1979), p. 534.

affectée : elle répartit la probabilité associée aux mondes-face sur les parties temporelles qui les habitent, elle fait de même pour les possibilités centrées des mondes-pile qui se trouvent être deux fois plus nombreuses et donc moins « pourvues ». Ainsi elle croit FL au degré 1/2, alors qu'elle croit PL et PM au degré 1/4 : ce sont les chiffres donnés par Lewis. Lorsque la Belle apprend ¬PM, elle revient à une situation d'équilibre par redistribution de la probabilité de PM sur PL, qui est alors crue au degré 1/2 comme FL. L'erreur serait ici de conditionaliser sur ¬PM comme le fait Lewis et de croire face à un degré supérieur. Si conditionalisation il y a, c'est une « conditionalisation compartimentée »[160]. Elle consiste exactement à ne distribuer la probabilité d'une possibilité centrée éliminée qu'entre les possibilités centrées du même monde non centré (à moins qu'il n'y en ait plus, auquel cas on se tourne vers les autres mondes, comme le ferait la conditionalisation standard).

Comme nous l'avons écrit au chapitre 2, Mikaël Cozic fait un lien entre la conditionalisation compartimentée, l'imaging et la mise à jour, c'est-à-dire le réajustement des probabilités qui a lieu quand un sujet apprend un *changement* dans son environnement, et non une information à propos d'un environnement supposé stable. Quand la Belle apprend qu'on est lundi, elle apprend comment a *changé*, depuis ses croyances initiales du dimanche, une caractéristique de sa situation. S'il s'agit bien d'un contexte de mise à jour, les règles de l'imaging (et celles de Meacham) devraient s'appliquer[161]. Joel Pust arrive lui aussi au compte double demiste par une voie un peu différente : il considère notamment que les informations

[160] Meacham n'entend pas exactement le mot « compartiment » comme nous l'entendons depuis le dernier chapitre, mais ce n'est pas grave : nous devons comprendre ici que la nouvelle conditionalisation ne concerne qu'un nombre réduit de compartiments (au sens indiqué au dernier chapitre).

[161] Cozic (2007).

temporelles reçues par la Belle sont inoffensives parce qu'elles sont des propositions à accessibilité limité (au sens de Perry)[162].

### 4.2. Critique du double demisme du bayésianisme adapté

Meacham et Cozic pensent qu'à côté de la conditionalisation classique, qui a prouvé son efficacité, doit exister une révision adaptée aux défis de l'auto-localisation. Ils partagent donc avec certains tiéristes une volonté de réforme, mais elle s'engage dans une voie différente. Leur nouvelle conditionalisation intervient lorsque l'agent acquiert des informations auto-localisantes, elle déplace des probabilités à la manière de la révision standard mais se restreint aux mondes qui abritent les possibilités centrées éliminées. Si toutes les possibilités centrées d'un monde sont éliminées, c'est un monde possible qui est éliminé, et dans ce cas les déplacements de probabilité ont lieu sur l'ensemble des mondes à la manière classique.

La conditionalisation classique fonctionne donc dans des cas particuliers que la nouvelle conditionalisation traite aussi : cette dernière pourrait être considérée comme une généralisation de la première. Attention : il reste des zones d'ombre. Soient trois mondes $m_1$, $m_2$ et $m_3$, deux centres $c_1$ et $c_2$, et six mondes centrés $(c_i, m_i)$ qui constituent six possibles conjointement exhaustifs qu'on suppose équiprobables *a priori*. Que se passe-t-il lorsque une information élimine simultanément $(c_1, m_1)$, $(c_2, m_1)$ et $(c_1, m_2)$ ? Est-ce que les trois mondes centrés restants sont *a posteriori* équiprobables parce que le monde $m_1$ a été éliminé ? Est-ce que $(c_2, m_2)$ devient plus probable que les deux autres parce qu'il concerne un

[162] Nous les avons évoquées dans notre critique du tiérisme bayésien. Pust (2012) est sur ce sujet un article très complet qui, comme tous les textes de cet auteur sur la Belle au bois dormant, laisse planer un doute sur le type de résolution qu'il adopte. Le double demisme se démarque malgré tout. Le tiérisme est un adversaire récurrent de Pust, qui toutefois l'a aussi défendu, sans y voir de contradiction.

monde qui n'a pas été complètement éliminé alors que les deux possibilités centrées de $m_3$ sont sauves ? Il y a un flou.

L'épistémologie de Meacham devient sujette à critiques quand le temps y joue un rôle crucial. Selon Bradley (2011b), certains tiéristes considèrent que toutes les informations temporelles peuvent modifier des croyances éternelles : c'est une conception extrême qui incite par exemple Arntzenius à croire que le prisonnier devient de plus en plus certain, avec le temps qui passe et sans recevoir d'autres informations, que la lampe de sa cellule restera allumée. Le double demisme tomberait dans l'autre extrême : jamais les informations temporelles ne peuvent modifier les croyances éternelles. Rappelons que Bradley distingue découverte et mutation, c'est pour lui une dichotomie beaucoup plus pertinente que simplement temporel/éternel, elle permet de ne pas tomber dans les extrêmes. Il estime que, quand la Belle apprend qu'on est lundi, elle *découvre* la vérité d'un contenu dont la valeur de vérité n'a pas changé, elle n'est pas dans le cas d'une mutation de croyances due à l'écoulement du temps (exemple de la montre de Meacham). Lewis a donc raison de conditionaliser à l'accoutumée, et lorsqu'il écrivait « Tout le reste est comme avant » en 1979, il pensait sans être explicite aux découvertes, évidemment pas aux simples mutations.

Le lien établi par Cozic entre la nouvelle conditionalisation, l'imaging et la « mise à jour » (au sens de Walliser et Zwirn) peut aussi être interrogé. Il ne fait guère de doute que la mise à jour est indiscernable de l'imaging. Nous avons vu aussi qu'il est très tentant de réviser par imaging les possibilités centrées du problème de la machine à dupliquer. Mais nous envisagions dans ce problème des individus au même temps *t*, qu'ils soient dans un même monde possible ou dans un autre : nous avions donc une énigme d'auto-localisation, certes, mais plus spécifiquement de recherche d'identité (« suis-je l'original ou la copie ? »), où le rôle du temps parait mince. Ces différences inquiètent quelques philosophes, trop peu en

réalité ; mais admettons que nous nous inquiétons pour rien. Il reste qu'il est difficile de penser que la Belle est dans un contexte de mise à jour lorsqu'elle apprend qu'on est lundi. Cozic lui-même doute.

### 4.3. Le double demisme du bayésianisme ouvert

Voici un double demisme très spécial mais très intéressant qui n'est défendu que dans des publications de Nick Bostrom[163]. Nous disons qu'il est bayésien dans le sens où il prétend que la théorie bayésienne n'est pas troublée par l'auto-localisation, qu'elle est encore très bien telle qu'elle est, qu'en son cœur la conditionalisation n'a besoin d'aucun complément, d'aucune généralisation ; nous disons qu'il est ouvert parce qu'il se permet de longues remarques d'inspiration fréquentiste qui toutefois ne sont pas essentielles pour goûter son argument-clé, ici résumé.

Et si nous commettions une maladresse en ne distinguant que trois compartiments associés à FL, PL et PM ? Bostrom propose de distinguer non pas trois parties (et contreparties) temporelles du sujet de l'expérience, mais cinq : deux dans le monde-face[164], à savoir la partie de la Belle qui, le lundi, ne sait pas dater ce jour (appelons-la *f-l*), et la partie qui sait qu'on est lundi ($f\text{-}l_+$) ; de même, trois dans le monde-pile, à savoir, là encore, la partie de la Belle qui, lundi, ne sait pas dater ce jour (*p-l*), la partie qui sait qu'on est lundi ($p\text{-}l_+$), et enfin la partie qui, mardi, ne sait pas dater ce jour (*p-m*). À chaque partie correspond une proposition centrée :

FL : « Ma partie temporelle actuelle est *f-l* »

[163] Et notamment Bostrom (2007).

[164] Bostrom parle du monde-face, du monde-pile au singulier, manière économique de parler de classes de mondes, ou plus vraisemblablement restriction des mondes possibles à deux sans explication, en garantissant tacitement que cela n'invalide pas l'argument. Du coup, les parties temporelles de l'agent sont elles aussi réduites en nombre.

$FL_+$ : « Ma partie temporelle actuelle est *f-l*$_+$ »

PL : « Ma partie temporelle actuelle est *p-l* »

$PL_+$ : « Ma partie temporelle actuelle est *p-l*$_+$ »

PM : « Ma partie temporelle actuelle est *p-m* »

Bostrom tient notre relation ($R_1$) pour acquise, car elle est pour lui plus intuitive que son alternative tiériste[165]. Nous pouvons notamment poser les probabilités suivantes :

$C_{rév}(FL) = 1/2$ $\quad\quad$ $C_{lun}(FL) = 0$

$C_{rév}(FL_+) = 0$ $\quad\quad$ $C_{lun}(FL_+) = 1/2$ ? ou 2/3 ?

Essayons de calculer $C_{lun}(FL_+)$ en appliquant comme d'habitude le théorème de Bayes après conditionalisation sur $(FL_+ \vee PL_+)$, qui est l'information reçue quand les expérimentateurs font leur annonce lundi :

$$C_{lun}(FL_+) = C_{rév}(FL_+ \mid FL_+ \vee PL_+)$$
$$= C_{rév}(FL_+ \vee PL_+ \mid FL_+).C_{rév}(FL_+) / C_{rév}(FL_+ \vee PL_+).$$

Puisque $C_{rév}(FL_+) = C_{rév}(FL_+ \vee PL_+) = 0$, nous sommes parvenus à une fraction avec 0 au numérateur comme au dénominateur, donc une expression indéfinie. De quelque façon qu'on s'y prenne, $C_{lun}(FL_+)$ ne peut pas être obtenue par un calcul utilisant le principe de conditionalisation. Mais la démarche logique chère aux bayésiens est tout à fait respectée par Bostrom, qui ne suggère surtout pas de la remplacer par une autre. Alors comment doit-on estimer rationnellement $C_{lun}(FL_+)$ et $C_{lun}(F)$ ? Par exemple par alignement sur des chances objectives, comme le proposent la plupart des tiéristes : la bonne estimation est 1/2, et elle est bien plus intuitive que le 2/3 demiste.

[165] Bostrom analyse plusieurs variantes de *Sleeping Beauty* pour parvenir à cette appréciation à laquelle n'adhèrent pas tous les chercheurs. Dans l'une d'elle, *Extreme Sleeping Beauty*, la Belle est réveillée un million de fois si la pièce tombe sur pile.

### 4.4. Critique du double demisme du bayésianisme ouvert

Nous pensons que Bostrom est rusé, incroyablement rusé. Kai Draper avait surpassé presque tout le monde avec son *cil* (« cet instant lundi »), il a maintenant trouvé son maître. Cependant nous croirons peut-être que le raisonnement du philosophe suédois est fallacieux. Quand la Belle au réveil sait seulement (FL $\vee$ PL $\vee$ PM), elle s'attend à recevoir, pour clarifier sa situation, une information telle que FL $\vee$ PL, et ce qu'elle obtient serait quelque chose comme $FL_+ \vee PL_+$ ! Pour le (simple) demiste Darren Bradley, le raisonnement revient à dire que, lorsque la Belle apprend qu'on est lundi, elle apprend aussi qu'elle vient d'apprendre qu'on est lundi, autrement dit qu'elle gagne une information centrée qui la renseigne davantage sur sa localisation temporelle. C'est une argutie qui cache mal le fait que Bostrom ne se sert finalement jamais d'une règle de révision doxastique, ni traditionnelle ni proposée pour l'occasion, ce qui lui permet de faire passer les estimations probabilistes qu'il souhaite, avant comme après l'annonce des expérimentateurs[166]. Le jugement est sévère mais inévitable. Le double demiste n'explique pas clairement combien il faut distinguer de compartiments pour résoudre les problèmes d'auto-localisation de ce type, ni pourquoi pour certains problèmes il n'est pas nécessaire d'en distinguer autant qu'ailleurs pour parvenir à une solution.

Une fois qu'on a dit ça, il faut examiner les choses de plus près. Bostrom a des intuitions qu'il veut partager, il ne les exprime pas totalement. Il y a un avant et un après l'annonce « on est lundi », du temps s'écoule, de ce temps qui emporte les croyances temporelles qui concernent l'agent au premier chef : FL était possible, FL n'est plus, place à $FL_+$. Il y a comme une mutation cachée derrière l'apparence de la découverte, pour le dire avec les mots de Bradley : des possibles deviennent impossibles et inversement. Bilan : une présence accrue de probabilités 0 et d'expressions

[166] Bradley (2007), p. 136-141.

indéfinies. Jusque-là, Bostrom partage avec les autres doubles demistes une intuition d'un passage de relai des croyances que la conditionalisation ne capture pas. Pourquoi faire une différence entre une Belle qui ne sait pas que c'est lundi et une Belle qui le sait ? demandera-t-on. Ce que la Belle recherche est sa localisation, elle envisage FL au sens donné au début du chapitre, « la pièce tombe sur face et aujourd'hui est lundi », et que ce soit avant ou après l'annonce elle continue à avoir un crédit sur ce FL-là. Pas exactement, a l'air de dire Bostrom. Avant la Belle ne savait pas, après elle sait, mais pas n'importe quel savoir ! Avant elle se réveillait dans une expérience, elle ne datait pas le jour ; après elle peut dater le jour, elle se focalise soudain sur des possibles plus spécifiques. C'est assez pour croire que la Belle d'après n'est plus la Belle d'avant, assez pour croire que les croyances de la Belle d'après ne sont pas fondées sur celles de la Belle d'avant, assez pour séparer des « parties temporelles », mais même cette désignation évoque trop de continuité, pas assez de rupture, ne rend pas bien compte du phénomène. Celui-ci est encore non dit ou mal dit. La méditation philosophique à son sujet prendra du temps.

Ajoutons une remarque. Le titre d'un premier manuscrit de l'article de Bostrom annonçait une « synthèse des vues » tiériste et demiste, ou en tout cas des points de vue d'Elga et de Lewis. L'auteur a changé cette expression en « modèle hybride » dans la publication finale. C'est sans doute plus juste. Horgan propose une synthèse (critiquable) du tiérisme parce qu'il organise et unifie les parties isolées intéressantes, Bostrom prend un peu chez les demistes, un peu chez les tiéristes, il s'appuie beaucoup sur son intuition dans ses choix, puis a des difficultés énormes pour coller le tout. Il sait qu'il apporte ou veut apporter dans son propos quelque chose que le demisme et le tiérisme n'ont pas du tout. Cet apport tient peut-être du génie mais on ne peut pas savoir parce que la seule chose qui soit claire est malheureusement que ce qu'on lit chez Bostrom sur ce point précis est à l'état d'ébauche.

### 4.5. La réparation double demiste du demisme et du tiérisme

Le double demisme avait de quoi devenir le « réaligneur » privilégié des résultats probabilistes ontiques des résolutions « objectivistes » du paradoxe et des résultats épistémiques des résolutions « subjectivistes ». En effet, ce qu'il prétend réconcilier, c'est le tiérisme, le primitif, celui qui avoue avoir pour guide une fréquence de réveils et une propension de la pièce de monnaie, et le demisme, le premier, bayésien, qui rejette les fréquences et raréfie les propensions. Et finalement ?

Alors, oui, le double demisme propose que la Belle réponde 1/2 à la première question et encore 1/2 au renouvellement de la question qui suit l'annonce « on est lundi », et oui, il juge que la première réponse est très bien défendue par le simple demisme bayésien et que la seconde est le résultat d'un très bon argument tiériste, *mais pas fréquentiste ni même propensionniste*. Seul le résultat tiériste $C_{rév}(F) = 1/3$ a une forte dette envers le fréquentisme ; les doubles demistes le rejettent. Globalement, le double demisme est bayésien. Il dira d'abord du résultat demiste $C_{lun}(F) = 2/3$ qu'il est contrintuitif et mal démontré, et seulement ensuite qu'il s'éloigne des chances objectives, parce qu'au fond ce n'est pas son problème, la démonstration qu'il propose foule d'autres contrées. Nous pouvons sentir que le double demisme est représenté par des bayésiens plutôt objectivistes, le demisme lewisien par des bayésiens plutôt subjectivistes. Nous n'affirmerons pas pour autant que les doubles demistes s'inquiètent vraiment de la scission ontique-épistémique.

Cela dit, Bostrom est *peut-être* l'exception. Il faut préciser qu'il connait des demistes ouverts au fréquentisme tels que Leslie et Franceschi. Déjà, l'idée d'une Belle qui est incapable de se repérer dans le temps au jour près et se situe plus facilement dans l'expérience datable se trouve à la fois chez Bostrom et les demistes ouverts, certes pas toujours explicitée. Ensuite, Bostrom (2007) estime comme Leslie que la répétition de

l'expérience leurre le tiérisme et que pour ce problème particulier la distinction entre scénario répété et scénario singulier original est cruciale, mais contre Franceschi il propose d'évaluer, dans un scénario répété, des probabilités de face au réveil comprises entre 1/3 et 1/2.

Le scénario d'un n-*fold Sleeping Beauty Problem* s'étend sur *n* semaines consécutives. Par exemple, dans un scénario *2-fold*, on lance la pièce dimanche soir, on réveille la Belle lundi et, si pile, mardi, puis on la laisse dormir le reste de la semaine ; on relance la pièce le dimanche suivant, on réveille la Belle lundi et, si pile, mardi ; enfin l'expérience se termine par exemple le mercredi. La Belle ne peut dater ni le jour ni même la semaine lors de son ou de ses réveils, à cause de la drogue qu'elle prend avant de se rendormir. Voici qu'elle se réveille dans l'expérience *2-fold*. Bostrom considère pour elle quatre mondes non centrés désignés par : issue du premier lancer - issue du deuxième lancer. En suivant son intuition de l'équiprobabilité des mondes et de l'équiprobabilité des réveils dans un monde donné, il calcule les probabilités de douze possibilités centrées qu'il range dans un tableau de ce genre :

| | Première semaine | | Deuxième semaine | |
|---|---|---|---|---|
| | Lundi | Mardi | Lundi | Mardi |
| Monde-face-face : | **1/8** | | **1/8** | |
| Monde-face-pile : | **1/12** | | 1/12 | 1/12 |
| Monde-pile-face : | 1/12 | 1/12 | **1/12** | |
| Monde-pile-pile : | 1/16 | 1/16 | 1/16 | 1/16 |

Que doit se dire la Belle ? Si la pièce est tombée sur face lors du dernier lancer en date, une de ces quatre propositions est vraie : « on est le lundi de la première semaine dans le monde-face-face », « on est le lundi de la deuxième semaine dans le monde-face-face », « on est le lundi de la première semaine dans le monde-face-pile », « on est le lundi de la

deuxième semaine dans le monde-pile-face ». La Belle doit donc croire que le dernier lancer a amené face à un degré égal à la somme des quatre probabilités en gras dans le tableau, soit 5/12. Selon Bostrom, si la Belle participait à une expérience *999-fold* par exemple, elle devrait croire face à un degré extrêmement proche de 1/3.

Ainsi le philosophe se livre à une lecture ontologique de probabilités peu éloignée de celle du tiérisme fréquentiste. Il retrouve le 1/3 dans un scénario géant, tout en continuant à défendre le résultat $C_{rév}(F) = 1/2$ du scénario original. C'est malin. Et c'est sûrement ce qui fait de lui le double demiste non seulement le plus impliqué dans la tentative de réconciliation du demisme et du tiérisme, mais aussi le plus préoccupé par le problème de l'écart ontique-épistémique.

## 5. La désambiguïsation

Le dernier type de résolution du paradoxe accepte les résultats ($R_1$), ($R_2$) et ($R_3$), remplace ($R_4$) $C_{rév}(FL) = C_{rév}(F)$ par $C_{rév}(FL) = 1/3$, et explique que les réponses 1/2 et 1/3 que la Belle peut formuler au réveil (quand elle ignore la date) sont les réponses à deux questions distinctes que l'énoncé du problème confond malencontreusement. Il y aurait une ambiguïté dans l'énoncé, comme il y en aurait une par exemple dans l'énoncé du célèbre paradoxe de Bertrand (si l'on en croit des solutions en vogue), ou dans les versions imprécises et maladroites du Monty Hall. Il ne faut pas confondre la désambiguïsation de la Belle au bois dormant avec la suspension d'un jugement, l'impossibilité de trancher une question précise. D'après Monton et Kierland (2005) par exemple, les deux réponses 1/2 et 1/3 sont acceptables ; cela ne veut pas dire qu'un parangon de rationalité les accepte toutes deux en tant que réponses à une question ambivalente, cela veut juste dire que l'analyse des auteurs a tenté de les départager et n'y est pas

parvenue[167]. Imaginer que F et FL puissent avoir des probabilités différentes au même moment de l'expérience est très difficile. Les désambiguïseurs sont rares et nous allons peut-être tous les citer dans le développement qui suit et qui distingue une approche ontologique et une approche épistémologique.

### 5.1. La désambiguïsation ontologique ou fréquentiste

Les différents acteurs d'un débat dominé par le bayésianisme disent tantôt 1/2, tantôt 1/3. Pour les désambiguïseurs ontologiques, deux probabilités de ce qui semble être le même événement ou la même entité sont évaluées : tiéristes et demistes (ou plutôt doubles demistes) ont tous raison à ceci près qu'ils n'ont pas vu qu'ils ne calculent pas la probabilité d'une même chose.

Selon le mathématicien et physicien Berry Groisman, l'approche épistémique des probabilités est inadéquate alors que le fréquentisme convient parfaitement pour dénouer le paradoxe, à condition de préciser, si cela manque dans l'énoncé, ce qu'il faut compter lors d'expériences répétées pour établir une fréquence relative et donc une probabilité ontique objective. Quand je dis que je suis dans un réveil-face, j'affirme principalement que la pièce est tombée sur face, un réveil-face étant par définition un événement qui fait suite à cet autre événement qu'est l'obtention de face. Il est clair que les lancers de pièce et les réveils sont liés. Néanmoins, compter les uns ne conduit pas au même résultat que compter les autres puisque pile est suivi par deux réveils au lieu d'un. Le demisme compte des lancers ou, ce qui revient au même, des expériences ou encore des séries de réveils, tandis que le tiérisme compte des réveils élémentaires, des réveils-face-lundi, réveils-pile-lundi, réveils-pile-mardi.

[167] Par ailleurs, l'analyse de Monton et Kierland conclut que la Belle doit répondre 1/2 et non 2/3 à la question subsidiaire. Elle est donc défavorable au simple demisme.

Le réel et son organisation sont donc mal observés et mal exprimés dans nos jugements lorsque, rassurés par le lien de cause à effet entre lancers et réveils, nous nous focalisons sur les uns plutôt que sur les autres[168].

Toujours à la recherche de modélisations, Paul Franceschi, qui a évolué du demisme vers une « *two-sided ontological solution* », reprend l'idée de Groisman et améliore le propos. Il présente une analogie avec une urne pour le moins insolite puisque celle-ci peut contenir des couples de boules « hyper-enchevêtrées », composés de boules qui peuvent être mélangées et choisies comme des boules normales mais qui ne peuvent pas être extraites de l'urne sans que la boule qui leur est associée et qui semblait pourtant indépendante ne soit tirée elle aussi. Deux boules hyper-enchevêtrées forment donc un même objet aux propriétés comparables à celles de certains objets quantiques. Voici l'expérience de l'« *hyper-entanglement urn* », élaborée afin d'arbitrer le match entre le simple demisme et le double demisme, *pas le tiérisme* :

> Une pièce équitable va être lancée. Si elle tombe sur face, l'expérimentateur placera dans l'urne une boule rouge normale. En revanche, si elle tombe sur pile, il placera dans l'urne un couple de boules hyper-enchevêtrées, composé d'une boule rouge et d'une boule verte indissociablement liées. L'expérimentateur ajoute que la salle sera plongée dans l'obscurité complète, et que vous serez alors dans l'incapacité totale de détecter la couleur des boules, et pas plus capable de savoir, quand vous aurez retiré une boule de l'urne, si c'est une boule normale ou une boule qui fait partie d'un couple de boules hyper-enchevêtrées. L'expérimentateur lance une pièce, et au moment où vous tirez une boule de l'urne, il vous demande d'évaluer la probabilité que la pièce soit tombée sur face.[169]

La réponse est 1/2 mais ce n'est pas important. Expliquons sommairement pourquoi cette expérience récuse les résultats du simple

[168] Groisman (2008).

[169] Franceschi (2010), p. 2.

demisme. L'erreur d'une pensée demiste serait de schématiser hâtivement la situation de la Belle par la situation d'un sujet tirant une boule colorée dans une urne qui en contient une ou deux, toutes normales. Mais c'est oublier un point de la structure du paradoxe, déterminant pour un demiste : les deux réveils-pile sont indissociables, ils forment un même *objet*. L'expérience aux boules hyper-enchevêtrées montre l'erreur. Par exemple, la probabilité naïve de tirer une boule verte « en tant que couleur » (d'être dans un réveil-mardi en tant que segment de temps) est :

$$\begin{aligned} p(\text{Vert}) &= p((\text{Vert} \wedge \text{Face}) \vee (\text{Vert} \wedge \text{Pile})) \\ &= p(\text{Vert} \wedge \text{Face}) + p(\text{Vert} \wedge \text{Pile}) \\ &= p(\text{Vert} \mid \text{Face}) \times p(\text{Face}) + p(\text{Vert} \mid \text{Pile}) \times p(\text{Pile}) \\ &= 0 \times 1/2 + 1/2 \times 1/2 = 1/4. \end{aligned}$$

Cela correspondrait au résultat demiste $C_{\text{rév}}(\text{PM}) = 1/4$. Mais la probabilité qui ne néglige rien est la probabilité de tirer une boule verte « en tant qu'*objet* » (d'être dans un réveil-mardi en tant qu'*objet*), et elle est égale à 1/2, c'est-à-dire la probabilité de pile, puisque si pile est venu, on a toutes les chances de tirer une boule verte en tant qu'objet (boules rouge et verte étant solidaires). L'erreur demiste est de considérer parfois des mondes centrés comme des segments de temps et d'ailleurs de paniquer lorsque le tiérisme propose de répéter l'expérience. Si l'on poursuit l'analogie avec l'urne d'hyper-enchevêtrement, les estimations du double demisme sont confirmées, pas celles du simple demisme. La probabilité de face sachant qu'on a tiré une boule rouge (sachant lundi) est 1/2. Si l'on croit que c'est 2/3, on confond encore couleur et objet (segment de temps et objet).

Les probabilités tiéristes sont elles aussi étayées par une variante de l'expérience de l'urne où, cette fois-ci, des boules en tant que *couleurs* et non objets sont considérées. Double demisme et tiérisme interprètent différemment certains concepts ambigus, et dans le débat la confusion

s'ensuit. Le double demisme a pourtant raison de *toujours* prendre en compte l'indissociabilité des mondes centrés colocalisés jusqu'à ne plus voir que des mondes non centrés ; le tiérisme a un autre point de vue respectable dans lequel l'indissociabilité perd son sens, il considère alors *toujours* des segments de temps. L'un voit F, l'autre FL. Le simple demisme fait un affreux mélange des genres. Un parangon de rationalité est désambiguïseur : il discerne la nature des possibles et attribue une probabilité double demiste ou tiériste selon le cas.

### 5.2. Critique de la désambiguïsation ontologique

Une forme de radicalité gêne dans ce discours. Le bayésianisme est jugé impuissant, les probabilités épistémiques sont évacuées. Pourquoi pas ? Mais la question canonique du problème est : « À quel degré devez-vous croire que la pièce est tombée sur face ? » Les partisans de cette solution ontologique feraient donc plus que relever une ambiguïté qui noue le paradoxe, ils semblent suggérer que raisonner avec des degrés et des révisions de croyances produit ce genre de paradoxe. La Belle au bois dormant décrédibiliserait la théorie bayésienne et montrerait l'efficacité du fréquentisme et de l'analyse de l'aléatoire par le moyen de probabilités ontiques, les seules réellement objectives.

Tout de même, pour Franceschi au moins, un degré de croyance n'est pas une absurdité, loin de là. Alors se repose la question : à quel degré la Belle doit-elle croire que la pièce est tombée sur face ? Tout désambiguïseur répondra d'abord qu'il y a deux questions mêlées dans la formulation : on demande peut-être à la Belle son estimation de la probabilité que la pièce tombe sur face à une date précise connue, la nuit de dimanche à lundi, c'est-à-dire la probabilité d'être dans un réveil-face ; ou bien on lui demande d'estimer la probabilité que la pièce est sur face, montre le côté face maintenant, c'est-à-dire la probabilité d'être dans un

réveil-face-lundi. Puis, pour montrer que ce n'est pas la même chose, le désambiguïseur *ontologique* évite un discours sur l'équivalence logique de propositions éternelle et temporelle. Nous avons vu que tout son raisonnement consiste à ramener à des événements aléatoires subtils des possibilités centrées. Nous ne sommes pas persuadés que Groisman, pourtant sûr de lui, y arrive facilement ; Franceschi est peut-être plus prudent et plus ingénieux, même s'il a besoin d'examiner plusieurs problèmes d'urnes et que sa modélisation est encore imparfaite[170]. Enfin, à ce raisonnement le désambiguïseur peut ajouter : une fois que la Belle a sondé le réel et séparé les objets qu'une mauvaise observation confond, une fois qu'elle a trouvé toutes les probabilités objectives des événements qui la concernent, elle aligne ses crédits dessus.

Selon Franceschi, la Belle qui ignore qu'on est lundi accorderait alors des crédits différents à F (1/2) et à FL (1/3) ; sachant qu'on est lundi, elle leur accorderait le même crédit 1/2. Le demisme n'était pas autorisé à poser : $C_{rév}(PL) = C_{rév}(PM) = 1/4$[171] ; c'est bien au degré 1/3 que la Belle doit croire ces possibilités centrées. Rien que pour cette remarque profonde, l'article du philosophe français est un bienfait ! On peut tout critiquer, éventuellement croire que la double réponse ontologique de la Belle au réveil est contournement et non dénouement du paradoxe, mais on ne peut pas passer outre cette remarque. $C_{rév}(PM) = 1/4$ est tout aussi faux que $C_{lun}(FL) = 2/3$ pour un désambiguïseur. Or ce 2/3 nous semble plus contrintuitif que ce 1/4 et la littérature non demiste l'a en effet toujours commenté négativement, à la différence du 1/4 qui semblait issu du calcul le plus simple : si je crois en face(-lundi) au degré 1/2, je crois en pile-lundi

[170] Cette modélisation est une évolution de l'analogie avec l'urne et les boules rouge et verte, décrite il y a peu de temps dans cette thèse et jugée mauvaise par Franceschi lui-même.

[171] C'est l'« erreur demiste » que Franceschi expliquait avec l'exemple de la boule verte « en tant que couleur » ou « en tant qu'objet ».

et aussi en pile-mardi au degré (1 – 1/2) / 2 (en supposant que je dois être indifférent). Franceschi dit : non, c'est tantôt trouver les réveils-pile indissociables (série datable de réveils dans le même monde), tantôt les trouver dissociables (réveils élémentaires, « segments de temps »). Nous pensons qu'il faut reconnaitre ce trait du demisme ouvert au fréquentisme (l'ancienne « école » de Franceschi) qu'on retrouve aussi chez le double demiste Bostrom : se localiser dans l'expérience datable et se localiser dans un jour non datable sont, si l'on peut dire, des manières d'orienter sa raison complètement différentes. Allons plus loin : *un vrai demiste, un double demiste, concentré sur des mondes non centrés, ne pense pas naturellement à probabiliser des PL ou des PM. Quand il le fait, il quitte son point de vue pour se placer sur celui du tiérisme, mais doit alors accepter les probabilités d'un autre espace de possibles, sinon il déraisonne.*

### 5.3. La désambiguïsation épistémologique ou bayésienne

Cette approche du problème n'est tentée que par le philosophe *Peter* Lewis dans un article de 2010, « *Credence and self-location* », qui conteste l'extension à des propositions auto-localisantes du principe d'équiprobabilité de propositions logiquement équivalentes. Tant que la Belle n'est pas certaine que son réveil est un réveil-lundi, elle n'a pas à donner nécessairement un même crédit à F et FL. Plus encore, de très bons indices mènent à $C_{rév}(F) = 1/2$ et $C_{rév}(FL) = 1/3$.

F si et seulement si FL ? Ce n'est pas vrai tout le temps. Dans le contexte de l'expérience, c'est vrai. Mais P ssi PL est faux au réveil puisque PM est une possibilité pas encore éliminée, et que P n'implique pas PL. Ces simples remarques nous incitent à la prudence. Un agent a souvent l'impression que « Je suis localisé quelque part » n'est pas un surplus d'information et ne modifiera pas une croyance éternelle. Mais au réveil la Belle envisage plus de positions temporelles possibles dans les mondes-pile

que dans les mondes-face. Il apparait alors que dire « Je suis dans un monde-face et localisé quelque part dans ce monde » est dire plus que « Je suis dans un monde-face ». Cette asymétrie entre les centres possibles en un monde et ceux possibles dans un autre monde fait que les crédits associés à F et à FL se désolidarisent.

D'habitude, l'argument du pari hollandais justifie la règle selon laquelle A ssi B implique C(A) = C(B) : si un agent sait que A ssi B et s'il assigne malgré tout $C(A) \neq C(B)$, il est possible de lui proposer un ensemble de paris sur A et B qu'il trouvera équitables mais qui le conduiront nécessairement à une perte s'il les accepte tous ; cela montre l'incohérence des deux probabilités. Mais la situation de la Belle au réveil est particulière. Elle sait que F ssi FL ; supposons qu'elle opte malgré tout pour $C_{rév}(F) = 1/2$ et $C_{rév}(FL) = 1/3$, des probabilités proposées par les désambiguïseurs ontologiques et qui correspondent aux intuitions demistes et tiéristes. Le problème est la place et le nombre des paris dans une expérience qui doit durer un ou deux jours et dont le sujet est drogué pour oublier ses réveils. En effet, lorsque les paris sont « synchroniques », c'est-à-dire proposés une fois, mettons le lundi, on peut objecter que le bookmaker n'est pas dans le même état épistémique que la Belle, il sait au moins quand est lundi et quand proposer le pari, donc il est plus facile pour lui de « l'escroquer », à moins qu'il partage son savoir avec elle, ce qui fausserait l'épreuve des paris[172]. À l'inverse, si les paris sont « diachroniques », c'est-à-dire proposés à chaque réveil par un bookmaker engagé dans l'expérience avec la Belle, réveillé avec elle, rendu amnésique avec elle, les deux personnes seront, certes, dans le même état épistémique, mais les paris seront déséquilibrés puisque proposés une fois en cas de face, deux fois en cas de pile[173].

[172] Cette objection est formulée par Hitchcock (2004).

[173] Cette objection est formulée par Bradley et Leitgeb (2006).

Mais en réalité le bookmaker n'a pas à être dans le même état épistémique que la Belle, les règles de l'expérience doivent simplement garantir que la Belle pourra déduire pertes ou profits sur un ensemble de paris qu'elle juge équitables. Bien sûr, elle ne doit pas en savoir trop, notamment elle ne doit pas savoir que c'est lundi qu'on lui proposera tel pari unique, sinon elle apprendra que c'est lundi quand le pari lui sera proposé. Ensuite, s'il est normal de considérer que des paris sur les mondes-F ne sont équilibrés que quand il est prévu de les proposer dans les mondes-P autant de fois que dans les mondes-F, c'est-à-dire en nombre fixe quel que soit ce (notre) monde, il faut aussi considérer que des paris sur les mondes-centrés-FL ne sont équilibrés que quand il est prévu de les proposer autant de fois à chaque réveil. Un système de paris équilibrés jugés équitables et acceptés par une Belle qui estime $C_{rév}(F) = 1/2$ et $C_{rév}(FL) = 1/3$ est par exemple : au réveil du lundi, 3 € sur F à 1 contre 1 ; à chaque réveil (lundi, puis mardi si pile), 4 € sur ¬FL à 1 contre 2 (2 € de gain si ¬FL, 4 € de perte si FL). À ce jeu, la Belle perd 1 € si face et gagne 1 € si pile : ce n'est pas une situation de pari hollandais. On voit facilement qu'on peut faire varier les mises autant qu'on veut (les cotes restant inchangées, pour l'équité) sans jamais parvenir à un pari hollandais. L'inférence de F ssi FL à $C_{rév}(F) = C_{rév}(FL)$ ne peut pas être justifiée par l'argument *Dutch book* dans un contexte Belle au bois dormant.

Cela ne veut pas dire que $C_{rév}(F) = 1/2$ et $C_{rév}(FL) = 1/3$ sont les probabilités que la Belle doit adopter, le raisonnement ci-dessus pouvant être tenu avec d'autres chiffres. Peter Lewis propose cependant des paris hollandais qui disqualifient des probabilités concurrentes. L'invulnérabilité aux paris hollandais de $C_{lun}(FL) = 1/2$ et $C_{lun}(F) = 1/2$, et la faillibilité de $C_{lun}(F) = 2/3$ peuvent aussi être montrées. Il est normal qu'une Belle qui, en apprenant qu'on est lundi, rééquilibre le nombre de ses positions temporelles possibles dans les mondes, rende équiprobables F et FL. Il est finalement rassurant de se retrouver avec $C_{lun}(F) = C_{rév}(F) = C_{dim}(F) = 1/2$ :

ces résultats ne menacent pas les principes de conditionalisation et de réflexion, ni le principe principal. Ces lois sont pensées pour l'espace logique classique, les difficultés et les amalgames surviennent avec l'auto-localisation, David Lewis lui-même en fait les frais.

### 5.4. Critique de la désambiguïsation épistémologique

Les solutions par désambiguïsation de la Belle au bois dormant, certes rares et proposées assez tard, ne sont pas très commentées dans la littérature. Nous pensons qu'avec Peter Lewis vient un argument audacieux, de belle allure, malheureusement miné de points contestables.

Son explication hâtive autour du surplus d'information dans « Je suis dans tel monde et localisé quelque part dans ce monde » est pour le moins peu convaincante. D'abord parce que Peter Lewis croit montrer (ou devrait avoir pour objectif de montrer) que des propositions équivalentes peuvent ne pas être équiprobables, alors qu'il montre plutôt que des propositions qui cessent d'être équivalentes ne sont plus équiprobables. Ensuite, il passe trop vite du général au particulier : dans une situation d'« asymétrie » des positions temporelles possibles, par exemple dans l'expérience de la Belle, « Je suis dans tel monde et localisé quelque part dans ce monde » *semble* en effet dire plus que « Je suis dans tel monde » ; ce n'est pas pour cela que « Je suis dans un monde-face et on est lundi » dit plus que « Je suis dans un monde-face ». Pourquoi faudrait-il attendre de savoir qu'on est lundi et donc de réparer l'asymétrie pour que FL ne dise rien de plus que F ? Pourquoi ne suffit-il pas de savoir qu'*il n'y a qu'une position temporelle possible dans les mondes-face* ?

Venons-en aux paris. Nous aussi sommes enclin à penser que la Belle doit pouvoir déduire pertes ou profits sur un ensemble de paris qu'elle juge équitables. Mais pour cela, apparemment, la Belle doit savoir dès dimanche combien de fois tel pari lui sera proposé, selon le cas en ce monde ou à

chaque position temporelle. Imaginons la Belle qui se réveille et se voit proposer un pari qu'elle sait être unique dans l'expérience : certes, elle ne sait toujours pas que c'est lundi puisqu'elle croit possible que le bookmaker propose le pari mardi en cas de pile, mais elle apprend qu'elle est réveillée le jour où on lui propose le pari unique, et il y avait en cas de pile possibilité qu'elle ne soit pas réveillée ce jour-là. Pour *certains* analystes du paradoxe qui penseront que cette information est bien pertinente pour l'issue pile/face, il y a de quoi modifier quelques probabilités. Voilà qui fausserait l'épreuve complexe des paris de Peter Lewis, qui propose notamment un pari unique sur F, à côté d'un pari diachronique sur FL. Et le choix du pari diachronique est peut-être lui-même contestable.

Comme ses deux collègues fréquentistes, le désambiguïseur bayésien essaie d'additionner les points de vue double demiste et tiériste. Il considère que l'un est une focalisation sur les croyances éternelles tandis que l'autre est une focalisation sur les croyances temporelles ou mixtes. Si l'asymétrie des compartiments du problème trouble les chercheurs, elle ne trouble pas un parangon de rationalité qui distingue une dynamique des croyances d'un genre et une dynamique des croyances de l'autre genre, cohérentes l'une envers l'autre mais pas au point de toujours conserver, par exemple, l'équiprobabilité de propositions des deux genres contextuellement équivalentes.

### 5.5. Désambiguïsation et désamour ontique-épistémique

Les désambiguïseurs ont un esprit de synthèse, mais limité à une seule approche des probabilités. Groisman et Franceschi réconcilient le fréquentisme des tiéristes primitifs avec un insoupçonné fréquentisme demiste, donc réconcilient le fréquentisme avec lui-même. Peter Lewis fait quelque chose de similaire avec les solutions bayésiennes. On peut se demander si c'est là la scission ontique-épistémique à son stade le plus

extrême et irréparable, ou bien au contraire si une synthèse supérieure des deux désambiguïsations serait l'alignement parfait.

Les désambiguïseurs ontologiques sacrifieraient volontiers la probabilité épistémique pour ne garder que l'ontique. Un tel sacrifice d'une partie sur deux, dira-t-on, est une manière de faire disparaitre une scission, puisqu'il n'y a scission qu'entre deux parties. C'est oublier que de l'autre côté Peter Lewis est toujours là, en gardien de la probabilité épistémique. Scission reconduite ? Pas tout à fait. Ce philosophe semble intéressé par l'interprétation ontique de la probabilité. Il a lu Groisman et a rédigé un rapide commentaire :

> Groisman plaide pour une conclusion similaire, à savoir qu'il y a deux propositions « face », auxquelles il faudrait assigner les probabilités 1/2 et 1/3 respectivement. Cependant, il ne fait pas la distinction en termes de propositions centrées et non centrées, et ne s'occupe pas de l'inférence de A ssi B à P(A) = P(B).[174]

Comme on le voit, le fréquentisme prononcé de Groisman n'effraie pas le bayésien, qui constate que le mathématicien arrive à des conclusions similaires par un autre cheminement intellectuel. Mais Lewis aurait peut-être été plus enthousiaste avec le discours de Franceschi, s'il en avait eu connaissance. Ce qui laisse penser cela est un article qu'il a écrit avant « *Credence and self-location* », dans lequel il ne trouve pas absurde que la Belle estime dimanche qu'elle sera réveillée lundi… avec probabilité 1/2 ! Il parvient à ce monstre en partant d'un parallèle entre une expérience où la Belle est simplement réveillée lundi et mardi dans tous les cas (et oublie ses réveils) et l'expérience everettienne de l'observateur quantique Q qui mesure par exemple le spin *x* d'un électron dans un état propre de spin *z* : Q se divise en deux observateurs $Q_{up}$ et $Q_{down}$, un qui voit le résultat « spin *up* », l'autre qui voit « spin *down* ». Q est psychologiquement continu avec

[174] P. Lewis (2010), note 15.

$Q_{up}$, et aussi avec $Q_{down}$, mais $Q_{up}$ n'est pas continu avec $Q_{down}$. De façon similaire, la Belle du dimanche est psychologiquement continue avec sa partie temporelle du lundi, mais aussi avec celle du mardi, car il n'y a pas continuité psychologique entre lundi et mardi à cause de l'amnésie. Or, les everettiens admettent bien souvent qu'avant sa ramification l'observateur Q doit estimer à 1/2 la probabilité de voir « spin *up* ». Si ce n'est pas l'analyse de la Belle au bois dormant qui doit inviter les everettiens à revoir cette probabilité mais bien l'inverse, les chercheurs devraient réfléchir à une Belle qui estime dimanche la probabilité qu'elle sera réveillée lundi à 1/2, et pas 1[175].

Cet article de Peter Lewis a fait réagir les philosophes David Papineau et Víctor Durà-Vilà ; nous pressentons avec raison qu'ils lui reprochent beaucoup de choses[176]. Ce n'est pas notre problème ici. Lewis fait appel à une analogie avec une expérience quantique qui n'est pas sans rappeler l'urne quantique de Franceschi. Toutefois, il apparait des différences : il n'est pas question d'enchevêtrement ni d'indissociabilité de deux objets, les probabilités sont interprétées différemment, les conclusions ne sont pas les mêmes. Et pourtant, à bien y regarder, ce Lewis qui précédait le Lewis désambiguïseur est très semblable au Franceschi qui précédait le Franceschi désambiguïseur : certes, il n'est pas un simple demiste revendiqué, mais il a conscience d'apporter des éléments en faveur du demisme ; en outre, sa probabilité est très objective et par moment ontique, c'est la probabilité qu'un sujet lit dans ce qui est assimilable à un processus aléatoire, et l'étonnant 1/2, degré de croyance du dimanche en un réveil lundi, est inexplicable sans ce fondement aléatoire. La Belle du lundi et la Belle du mardi ne sont pas continues, mais comme présentes en des mondes différents. Tout se passe comme si le monde se scindait en deux

[175] P. Lewis (2007).

[176] Cf. Papineau et Durà-Vilà (2009a, 2009b) et P. Lewis (2009).

possibles, un monde-lundi et un monde-mardi, juste le temps d'une expérience. Par conséquent, l'expérience originale de la Belle doit être traitée comme la variante qui nous apparaissait comme un travers demiste dans lequel Franceschi, justement, semblait jadis tomber :

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **Face** | 1 | 0 |
| **Pile** | 1 | |

L'auteur écrit enfin que quiconque accepte l'interprétation des mondes multiples doit être demiste, ce que contestent aussi Papineau et Durà-Vilà. Lewis démontre que les deux visages de la probabilité lui sont chers. Il n'est sûrement pas un subjectiviste radical qui aggraverait la scission ontique-épistémique, il aurait plutôt tendance à sentir ses thèses téméraires et ses probabilités épistémiques, indéfendables au premier abord, dans un monde de l'aléatoire complexe. Finalement, l'avenir des deux désambiguïsations, l'ontologique et l'épistémologique, est peut-être bien une union de ces deux opposés, qui corrigerait leurs défauts évidents et les rendrait convaincants, une union qui ferait s'entendre sur leurs probabilités communes les fréquentistes et les bayésiens et effacerait entre eux toute méfiance.

# Chapitre 5

# Dénouer un paradoxe probabiliste

Nous proposerons bientôt, dans une ultime étape, une solution au problème de la Belle au bois dormant. Elle est certainement plus audacieuse et plus incroyable que les précédentes solutions et ses nouveaux concepts, tels le *basculement* et le *déphasage*, feront réagir même les personnes les plus capables de trouver un intérêt à toute argumentation suspecte. Elle ne manque pourtant pas de qualités. C'est une synthèse originale d'idées venues des quatre principaux positionnements, même si certains sont préférés à d'autres (nous nous en expliquerons). Surtout, elle prétend dénouer le « vrai » paradoxe, celui de la mésentente anormale de deux interprétations de la probabilité.

Nous partons de très loin avant de pénétrer le cœur de la résolution. L'amélioration de plusieurs variantes de la Belle au bois dormant trouvées dans la littérature, l'élaboration de variantes nouvelles, mais aussi l'examen de deux problèmes du raisonnement anthropique, vont être nécessaires pour approcher la révélation finale par une route bien éclairée.

# 1. L'art de varier la Belle au bois dormant

## 1.1. Retour sur la Belle et d'Alembert

Qui peut croire qu'intégrer au scénario de la Belle un deuxième lancer de la pièce de monnaie le mardi, qui plus est un lancer dont l'issue n'a pas de conséquence sur le déroulement de l'expérience, engendre de très importantes réflexions ? Pourtant il faut y croire. Dans le chapitre 1, nous avons remarqué que ce scénario alternatif ressemblait au jeu du croix ou pile en deux coups, mais placé du point de vue d'un observateur qui subit un incident cognitif qui déforme la perception du temps ; nous avons tenté de trouver grâce à ce scénario une explication aux probabilités de prime abord extravagantes données par D'Alembert. Michael Titelbaum, l'auteur de cette variante, a une autre analyse peut-être plus impressionnante puisqu'elle tente de montrer l'inconsistance des demismes simples et doubles en un calcul et quelques questions pertinentes[177]. Titelbaum, qui connait le demisme bayésien et les principaux auteurs doubles demistes, estime avec raison que le double demisme trouve l'idée d'une conservation de la croyance en face tout au long de l'expérience, malgré des apports de données auto-localisantes, bien plus défendable que le résultat contrintuitif $C_{lun}(F) = 2/3$ du simple demisme, et croit échapper, en abaissant cette probabilité à 1/2, à la critique tiériste des chances objectives de la pièce. Mais il n'y échappe pas.

Titelbaum expose une version non d'alembertienne du scénario alternatif, dans laquelle l'expérimentateur, qui n'est pas D'Alembert, ne voit aucun mal à relancer la pièce le mardi dans tous les cas, même si la Belle dort ce jour-là. Nous pensons effectivement que cela a son importance : le second tirage doit avoir lieu quel que soit l'issue du premier, pour le bien de la démonstration du philosophe. Reprenons depuis

[177] Titelbaum (2012).

le début : la Belle accepte de participer à une expérience dont elle connait toutes les règles. Le lundi, on la réveille pour un entretien, puis on la rendort avec la drogue à effet amnésique et on lance une pièce équilibrée. Si face, on la laisse dormir mardi, sinon on la réveille pour un nouvel entretien et on la rendort avec la même drogue. *Dans tous les cas*, on relance la pièce le mardi soir quand la Belle est endormie ; le résultat de ce second tirage est sans conséquence.

L'entretien débute. On demande à la Belle le crédit qu'elle accorde à l'hypothèse $F_{lun}$ selon laquelle aujourd'hui est lundi et la pièce lancée lundi tombe sur face. La Belle, qui vient de lire un long texte double demiste qui l'a beaucoup impressionnée, répond 1/2 car elle reconnait qu'on lui demande la probabilité de FL et donc de F. On lui demande si elle accorde un crédit non nul à l'hypothèse $F_{mar}$ selon laquelle aujourd'hui est mardi et la pièce lancée mardi tombe sur face. Elle répond par l'affirmative puisque $F_{mar}$ est clairement une possibilité. On lui fait alors remarquer cette équivalence : la pièce lancée aujourd'hui tombe sur face si et seulement si $F_{lun} \vee F_{mar}$. La Belle acquiesce. On lui demande enfin si la probabilité que la pièce lancée aujourd'hui tombe sur face est supérieure à la probabilité de $F_{lun}$ qu'elle a estimée au début, à savoir 1/2. Embarrassée, elle devrait répondre que c'est exact mais s'y refuse. Le calcul est très simple : 1/2 plus un réel positif non nul, ça fait plus que 1/2. Pourtant elle sait que les expérimentateurs peuvent mettre dans sa main la pièce qui doit être lancée ce soir, et même lui demander de la lancer pour eux. La propension de la pièce à tomber sur face est 1/2, pas plus, pas moins. Déconcertée, la Belle doit revoir sa première estimation concernant $F_{lun}$. La probabilité tiériste 1/3 semble tout à coup plus appropriée.

Un simple calcul, pas de conditionalisation… Oui, le double demisme mais aussi le simple demisme ont un gros souci avec cette variante et son examen. Néanmoins, Titelbaum sous-estime la pensée du double demiste Nick Bostrom et surtout n'a pas du tout en tête les analyses des

désambiguïseurs qui voient le double demisme comme une des solutions du problème (une fois que l'ambiguïté de l'énoncé est effacée). L'indexicalité essentielle de « la pièce lancée aujourd'hui » ou encore de « aujourd'hui est lundi » ne fait aucun doute si la Belle ignore la date du jour. Titelbaum demande à la Belle de se représenter ses parties temporelles, d'estimer les probabilités de mondes centrés. Or, se localiser dans des mondes possibles mais pas dans le temps, compter des expériences et non des réveils, est crucial dans l'appréciation des doubles demismes ouverts au fréquentisme. La variante de Titelbaum ne peut déconcerter que des demistes et doubles demistes bayésiens et il se trouve que tous ont en commun de ne pas être très préoccupés par la scission ontique-épistémique. Mais les Bostrom, Groisman, Franceschi reprocheront à Titelbaum de ne pas avoir compris le « vrai » double demisme ; Peter Lewis en particulier continuera à soutenir que $C_{rév}(F) = C_{lun}(F) = 1/2$ ne contredit pas $C_{rév}(F_{lun}) = 1/3$. Il se trouve que ces chercheurs s'inquiètent de la scission ontique-épistémique mais y répondent à leur manière, parfois radicale. Nous pensons que ce n'est pas un hasard.

### 1.2. Vers un nouveau type de résolution

Deux options s'offrent à nous : écarter ces avocats du « vrai » double demisme, en pariant sur l'inconsistance de leurs thèses, en cédant à l'idée que leurs résultats probabilistes qui défient la logique ne sont pas la solution d'un paradoxe mais au contraire sa fixation ; ou bien les garder, leur trouver une place à côté du tiérisme. Car, quoi qu'il en soit, il faut bien avouer que le tiérisme sous ses multiples formes ou étapes ne peut pas être pris à la légère : il a engagé en conscience la « réparation » qui nous tient à cœur, a peut-être réussi à expliquer une probabilité de face de 1/3 (au réveil) à la fois ontologiquement et épistémologiquement ; une majorité de chercheurs y adhèrent, sans compter les désambiguïseurs qui lui font une place de choix, à côté du double demisme justement. Tout cela montre que

le tiérisme repose sur une intuition et une compréhension du problème fortes.

Quiconque a passé beaucoup de temps sur l'énigme de la Belle au bois dormant, quiconque a oscillé entre le demisme et le tiérisme, ne peut nier que les désambiguïseurs ont eux aussi une certaine vision du nœud paradoxal qui devrait inspirer plus de monde, mais ils l'expriment par des formules logiques et des énoncés de probabilité qui manifestement ne sont pas ceux de tout le monde, n'inspirent pas confiance. Au moins, ils soutiennent des résultats qui s'expliquent ontologiquement et épistémologiquement, quoique Peter Lewis soit à peu près le seul à en prendre conscience, à un degré que nous ignorons.

Au contraire, le résultat $C_{lun}(F) = 2/3$ du simple demisme, rarement soutenu par une argumentation dans les textes demistes alors qu'il devrait nourrir un souci constant, n'a rien pour lui. D'une part, il ne reflète en soi aucune profonde « vision », il est simplement le résultat d'une conditionalisation sur l'information centrée FL $\vee$ PL *dont la justification se trouve aussi dans le tiérisme*, mais celui-ci part d'une probabilité *a priori* différente. On peut même être plutôt d'accord avec les analyses des croyances dynamiques et du passage du temps d'un chercheur demiste comme Darren Bradley, et refuser dans le même temps que sa Belle entre en contexte d'asymétrie ou de déséquilibre des centres possibles entre deux classes de mondes en ne changeant rien à ses croyances éternelles (inertie) et revienne en situation normale en conditionalisant et en changeant. D'autre part, on peine à trouver un support ontique à $C_{lun}(F) = 2/3$. Nous avons vu qu'on trouve *dans des variantes ou des modélisations* une probabilité ontique (fréquentielle généralement) de 2/3 qu'on peut « faire correspondre » à $C_{lun}(F) = 2/3$, mais les chercheurs insistent sur le fait que l'on se trompe si l'on infère de celle-ci le résultat demiste. Nous ne disons pas que David Lewis est dans l'erreur, nous n'affirmons pas sans nuance que seul le bayésianisme objectif, constamment attentif aux proportions des

objets du monde, voit clair. Nous disons en revanche que ce demisme part dans une direction subjectiviste qui n'est pas la nôtre. Il y a du D'Alembert chez Bradley, mais de ce D'Alembert qui n'inspire pas confiance, celui de la pensée magique. Bradley ne croit probablement pas que les boules de cristal existent en ce monde. Il évoque un autre monde dans lequel existent de telles choses qui peuvent prédire, par exemple, l'issue d'un tirage à pile ou face imminent. Très bien. Mais il explique ensuite que certaines informations temporelles sont des informations sur le futur, ou peut-être pas tout à fait, mais assez pour tordre la probabilité qu'on associerait d'habitude à telle issue d'un futur tirage aléatoire. Une boule de cristal en moins puissant serait dans notre monde. La Belle de Bradley pourrait inventer un moyen de gagner aux jeux de hasard en améliorant le protocole (cérémonial) de l'expérience ! Qu'attendons-nous pour prendre de la drogue à effet amnésique ?

Je me permets de raconter à la première personne une anecdote qui montre qu'il faut être prudent avec les paradoxes, probabilistes en particulier. J'ai réfléchi il y a quelques années au paradoxe de la Chambre d'exécution (*Shooting-Room Paradox*)[178]. Je me souviens avoir programmé une simulation informatique dans laquelle, suivant le protocole du scénario, je faisais en quelque sorte dépendre du tirage pseudo-aléatoire de l'ordinateur le sort de groupes de plus en plus grands d'individus d'une vaste population. Et j'ai cru pendant deux ou trois heures que si je marquais un individu dans la population, s'il se retrouvait dans un groupe, si le programme s'arrêtait avant le jet virtuel des dés et attendait que je presse une touche du clavier, je ferais alors l'expérience de ma vie : le processus

[178] Leslie (1992) décrit le paradoxe : vous êtes poussé dans une salle alors que vous savez que 90 % de ceux qui y sont entrés seront tués ; vous apprenez pourtant que vous sortirez vivant à moins que le prochain jet de deux dés n'amène un double six. En fait, les groupes de personnes successivement poussés dans la salle sont de plus en plus grands, de manière exponentielle, de sorte que la prédiction « 90 % seront tués » sera réalisée quand un double six sera enfin tiré.

pseudo-aléatoire serait comme biaisé très nettement en défaveur du groupe, qui serait en toute probabilité « exécuté », alors que les chances objectives sont d'une sur trente-six. Deux ou trois heures de déraison, de superstition, de quatrième dimension, chez quelqu'un supposément protégé par sa culture et son goût pour la philosophie et les mathématiques. Mais nous devons tous être lucides et peut-être même nous souvenir de quelques événements de nos vies refoulés par honte : chez les êtres humains, de tels moments de faiblesse sont fréquents. Croire le contraire ou se croire immunisé n'est pas une bonne attitude. Bien sûr, il est parfois facile de se ressaisir, de revenir assez vite d'un état de faiblesse tant il nous donne à croire l'incroyable, tant il est dissonant et suspect, tant il est rejeté par ce que nous sommes, croyons, savons, pensons habituellement. Mais il arrive peut-être qu'il s'installe durablement sans que l'on en prenne conscience, quand il n'est *pas très* dissonant.

D'Alembert n'est pas un imbécile, c'est un homme, un grand homme sûrement, mais pas plus qu'un homme. Nous dirons des simples demistes la même chose, ce n'est pas leur naissance au 20$^{e}$ siècle qui les immunise contre une superstition sans gravité qu'ils ne remarqueraient pas et que d'ailleurs les autres chercheurs ne remarqueraient pas, tant l'ouverture d'esprit est nécessaire quand on débat autour d'un paradoxe aussi puissant. Encore une fois, cela ne signifie pas que $C_{lun}(F) = 2/3$ est assurément une telle superstition, mais nous avons le droit de le suspecter quand nous avons des indices pour cela, quand nos efforts pour penser cette probabilité demiste échouent depuis des années. Seul celui qui n'a pas conscience de sa propre condition faillible ne comprend pas cela et fait d'une éternelle recherche de sens le seul moyen d'être juste avec la pensée d'autrui ; pendant ce temps, il ne peut pas s'engager sur une voie de résolution du problème.

Notre voie consiste à abandonner ce 2/3 demiste, à tenir à bonne distance le $C_{rév}(F) \neq C_{rév}(FL)$ des désambiguïseurs, moins suspect selon

notre analyse mais très suspect aux yeux de beaucoup de chercheurs, et à faire quelque chose avec tout le reste, y compris avec les intuitions, selon nous mal exploitées, des désambiguïseurs et de Bostrom notamment. Il nous faudra montrer pourquoi la Belle doit répondre 1/2 à la question subsidiaire du lundi, puisqu'il s'agit de la seule alternative au 2/3. Il nous faudra avant tout arranger ou bien départager les réponses possibles 1/2 et 1/3 de la question principale, en continuant à explorer des variantes du problème.

### 1.3. La Belle et le Prince

Voici une variante favorable au tiérisme que la littérature évoque beaucoup depuis 2006[179]. La Belle s'endort dimanche soir avec son Prince. Elle est réveillée, en compagnie du Prince avec lequel elle peut alors s'entretenir, selon le protocole de l'expérience originale (lundi si face, lundi et mardi si pile). Comme la Belle, le Prince est endormi à l'aide de la drogue à effet amnésique ; il ignore de quel côté est tombée la pièce et ne peut dater le jour de son réveil que si on lui donne l'information ; il connait le protocole entier. La différence est qu'il est réveillé lundi et mardi quel que soit le résultat du lancer de la pièce ; il peut donc arriver qu'il soit réveillé alors que la Belle reste endormie. Voici que le Prince se réveille et comprend que l'expérience est en cours. Respectueux du principe d'indifférence, il attribue la probabilité épistémique *a priori* 1/4 à chacune des quatre hypothèses centrées, exclusives et conjointement exhaustives :

FL : « La pièce tombe sur face et aujourd'hui est lundi »

FM : « La pièce tombe sur face et aujourd'hui est mardi »

[179] Neal (2006), p. 15-17 ; Stalnaker (2008), p. 63 ; Weatherson (2011) ; Delabre et Gerville-Réache (2015). Stalnaker et Weatherson remplacent le Prince par la Laide (Ugly).

PL : « La pièce tombe sur pile et aujourd'hui est lundi »

PM : « La pièce tombe sur pile et aujourd'hui est mardi »

Il découvre alors que la Belle est aussi réveillée. Bien qu'il sût que cette situation arriverait durant l'expérience, il reçoit une nouvelle information indexicale : ¬FM, autrement dit FL ∨ PL ∨ PM. La probabilité de FM est maintenant annulée et, par conditionalisation équivalente à un partage de l'ancienne probabilité 1/4 de FM entre les trois hypothèses restantes, le Prince doit maintenant attribuer à chacune des trois la probabilité 1/3, ne voyant pas pourquoi en favoriser une au détriment des autres. Par conséquent, il doit à présent croire au degré 1/3 que la pièce est tombée sur face. Cette partie du raisonnement tiériste ne plairait pas aux désambiguïseurs pour des raisons déjà évoquées, mais nous laissons la désambiguïsation de côté pour l'instant. Les demistes et *certains* doubles demistes l'acceptent, mais ils soulignent, à la manière de leur critique d'une variante similaire évoquée dans le dernier chapitre[180], que la Belle, contrairement au Prince, ne peut jamais avoir l'occasion d'apprendre que celui des deux sujets rationnels qui dans cette expérience peut ne pas être réveillé mardi, à savoir elle-même, est ou n'est pas réveillé, puisqu'un tel apprentissage nécessite la conscience, l'état de veille, donc l'événement dont la réalisation ou la non-réalisation, justement, est à apprendre. En d'autres termes, selon le demisme, comprendre que « j'apprends que je suis réveillé » est aussi absurde que « j'apprends que je dors », c'est comprendre que la dynamique des croyances de la Belle est différente de celle des croyances du Prince.

Poursuivons le raisonnement tiériste. La Belle et son Prince peuvent à présent discuter et surtout échanger leurs estimations de la probabilité de face. Si la Belle est tiériste, les estimations se rejoignent, ce qui ne produit chez elle aucun questionnement particulier. Si elle est demiste, les

[180] C'est la variante de Dorr (2002), dont s'est servi Terry Horgan.

estimations divergent : ce désaccord de deux agents rationnels n'est pas rare, mais il serait bien étrange qu'ils ne parviennent pas à faire concorder leurs estimations après avoir mis en commun leurs connaissances. Et si elle hésite entre 1/2 et 1/3 ? Ne doit-elle pas repenser au raisonnement tiériste du Prince, ne peut-elle pas adopter sa conclusion ? Si nous admettons que la Belle doit calquer son degré de croyance sur celui de son compagnon, nous pouvons encore objecter que la présence de celui-ci modifie de façon significative l'expérience originale. Alors, par variations successives, plaçons-nous progressivement dans les conditions de l'expérience originale, d'abord en supposant que le couple est séparé par un rideau, puis que le Prince se situe dans une chambre isolée sans aucun moyen de dialoguer avec sa bien-aimée ; finalement, examinons le cas où le compagnon n'est plus du tout là et où la Belle ne peut que l'imaginer et comprendre le raisonnement tiériste qu'il pourrait tenir. À quel moment devrait-elle cesser de croire au degré 1/3 en l'obtention de face ?

Robert Stalnaker approche le problème de la Belle au bois dormant par l'eccéitisme qui le caractérise : les mondes centrés sont assimilés par ce théoricien de l'auto-localisation à des mondes non centrés dont il soupçonne l'équiprobabilité ; il n'a aucun mal à trouver convaincante la variante de la Belle et du Prince[181]. Une divergence persistante des estimations probabilistes de deux agents rationnels, portant sur la même croyance au même moment, n'est pourtant pas une aberration dans d'autres situations (certes complexes). Imaginons une vaste population d'hommes et de femmes, chacun invité une fois dans sa vie à participer à une expérience de grande envergure conduite depuis longtemps par de puissants Organisateurs. Les « cobayes », qui ont pris connaissance de toutes les règles qui suivent, sont endormis profondément. Les Organisateurs s'arrangent pour réveiller, dans une chambre où ils ont préalablement caché

[181] Cf. Stalnaker (2008), chap. 3. L'analyse de la variante commence à la page 63. La fin du chapitre 2 de notre thèse évoque amplement l'eccéitisme de cet auteur.

sous un foulard une pièce de monnaie, chaque individu à trois reprises et toujours en compagnie d'un individu (chaque fois différent) du sexe opposé, avant de les rendormir avec la drogue à effet amnésique. De plus, chaque femme est réveillée deux fois dans une chambre où la pièce est orientée côté pile et une fois dans une chambre où elle est orientée côté face ; chaque homme, à l'inverse, est réveillé dans une chambre-pile une fois, dans une chambre-face deux fois. Les agents sont libérés après leurs trois réveils. Voici qu'Alice et Bruno, deux participants rationnels, se réveillent dans une chambre. À cause de la drogue, c'est pour eux comme le premier réveil. Ils voient le foulard. Alice croit d'abord au degré 1/3 qu'en dessous la pièce est orientée côté face, mais Bruno croit au degré 2/3 cette éventualité. Conflit ! Doivent-ils rester sur ces estimations ou doivent-ils par exemple se mettre d'accord sur une probabilité 1/2, si celle-ci a toutefois une justification épistémologique ou ontologique ? Difficile à dire.

On objectera que le cas de la Belle et du Prince est différent, parce que le Prince sait plus, ou a plus d'informations que la Belle ; c'est donc elle qui doit adopter l'estimation de son compagnon. Vraiment ? Le Prince a appris, certes, alors que la Belle n'a rien appris. Mais il a appris qu'il est réveillé avec la Belle, il a appris ¬FM… ce que la Belle savait déjà, en tout cas si l'on en croit certaines analyses du problème. On peut donc contester que l'estimation du Prince fasse autorité. Les cheminements mentaux des deux « compagnons » après leur réveil sont trop différents. Néanmoins, mettre en relation la Belle avec un compagnon est une bonne idée, il faut perfectionner l'argument en le faisant reposer sur une variante plus ingénieuse.

### 1.4. La variante des Quatre Belles

Nous avons proposé dans un article récent un problème où les compagnons sont au nombre de quatre :

> Quatre Belles participent à une expérience dont elles connaissent toutes les règles. Le dimanche soir, elles s'endorment. On tire alors au sort une des quatre, équitablement : elle est ainsi désignée pour être réveillée lundi et seulement lundi. Puis on tire au sort une Belle parmi les trois restantes : elle est désignée pour être réveillée mardi et seulement mardi. Les deux sujets restants seront réveillés lundi et mardi. Le lundi, on laisse dormir profondément la Belle désignée pour ne pas être réveillée ce jour-là, on réveille les trois autres, on les réunit pour un entretien, puis on les rendort après leur avoir administré une drogue qui leur fait oublier tout ce qui s'est passé dans la journée. Le mardi, on réveille les trois Belles qui doivent être réveillées ce jour-là, on les réunit pour un entretien, on les rendort avec la même drogue. Les Belles n'ont aucun moyen de savoir si on est lundi ou mardi, ni de connaitre le résultat précis des tirages au sort, sauf si on les renseigne dans la conversation (néanmoins, elles savent évidemment que la Belle qui dort a été désignée par le sort). […]
>
> Suivons Aurore, une des Belles engagées dans l'expérience. Elle se réveille avec deux autres Belles, toutes trois sont incapables de se repérer dans le temps au jour près ; la quatrième Belle continue à dormir. À quel degré Aurore doit-elle croire qu'elle est une « désignée », c'est-à-dire que le hasard, au premier ou au second tirage, l'a désignée pour qu'elle ne soit réveillée qu'un jour sur les deux jours de l'expérience, autrement dit juste aujourd'hui ?[182]

Tout se passe comme si Aurore participait à l'expérience résumée dans le tableau qui suit, mais en compagnie de trois amies, chacune suivant le programme d'une des quatre lignes de tableau que le sort lui a réservée :

[182] Delabre et Gerville-Réache (2015), p. 264.

| Jour / Tirage | **Lundi** | **Mardi** |
|---|---|---|
| **Désignée 1** | 1 | 0 |
| **Désignée 2** | 0 | 1 |
| **Non-désignée 1** | 1 | 1 |
| **Non-désignée 2** | 1 | 1 |

Nous pensons qu'Aurore ne reçoit pas à son réveil d'information pertinente *au sens demiste* pour son statut désignée/non-désignée. En effet, elle savait déjà qu'elle serait réveillée durant l'expérience en même temps que deux autres Belles ; disons même, pour aller au bout de l'idée, qu'elle savait qu'elle se verrait en état d'éveil, que la Belle restée endormie ne serait pas elle. Par conséquent, que la Belle endormie soit Lucie, Estelle ou bien Norah n'a aucune importance, cette information n'a de force que pour modifier les croyances sur le statut de ces autres Belles. En revanche, il est possible qu'un tiériste la trouve pertinente pour le statut d'Aurore[183].

Ce scénario a plusieurs intérêts. Premièrement, Aurore est engagée dans une expérience qui ressemble à l'originale de la Belle au bois dormant à tel point que, si elle pouvait ignorer la présence de ses amies, elle devrait répondre la même probabilité que la Belle originale à la question principale de l'entretien (qui porte sur l'issue face d'un tirage, rappelons-le), après avoir tenu des raisonnements identiques. Pour faire court : un tiériste simulerait la répétition de l'expérience et déterminerait la proportion des réveils-désignée parmi tous les réveils, ou encore rendrait équiprobables les six compartiments naturellement envisagés (correspondants aux cases où la

[183] Léo Gerville-Réache évoquait ce point dans une correspondance privée : il est possible que la connaissance de l'identité de la Belle endormie fasse basculer dans un point de vue tiériste tout agent réticent.

probabilité 1 est portée) et déterminerait la même proportion, à savoir 1/3, ou encore utiliserait une méthode bayésienne plus complexe (perte/gain d'information temporelle, conditionalisation adaptée…), mais dans tous les cas il parviendrait à une probabilité de 1/3 « d'être une désignée » ; un demiste tiendrait des raisonnements et des comptes similaires mais n'envisagerait pas les mêmes objets ou ne leur attribuerait pas une même probabilité (équiprobabilité des mondes possibles, mais pas des mondes centrés), et son argument bayésien basé sur l'inertie doxastique le mènerait à une probabilité de 1/2 « d'être une désignée », la même probabilité que le dimanche juste avant l'expérience. Attention : si Aurore venait à apprendre qu'on est lundi, elle ne devrait peut-être pas croire qu'elle est une désignée au degré 1/2 (réponse de la Belle originale à la question subsidiaire). Cela s'explique par le fait qu'Aurore, avant d'avoir cette information, tient pour possible qu'aujourd'hui soit mardi et qu'elle soit une désignée, alors que face-mardi est une impossibilité pour la Belle originale.

Deuxièmement, en étant accompagnée de ses amies, Aurore est en quelque sorte en mesure de discuter avec des représentantes de ses contreparties dans les autres classes de mondes possibles, comme si l'ambassadrice des mondes-désignée1, celle des mondes-désignée2, celle des mondes-non-désignée1, celle des mondes-non-désignée2 étaient réunies en un même monde en même temps par on-ne-sait-quel moyen mystérieux, l'une des quatre étant endormie comme le veut le protocole. En outre, la discussion, échange d'informations, pourrait être pour nous très instructive. En effet, les Belles sont à égalité face au protocole, autrement dit, même si les expérimentateurs ont prévu pour elles des programmes différents (l'une n'est réveillée que le lundi, une autre est réveillée lundi et mardi, etc.), les Belles réveillées ne savent pas qui suit quel programme, un hasard équitable en ayant décidé dimanche soir ; elles ne sont donc pas dans la situation de la Belle et du Prince qui connaissent parfaitement leurs programmes respectifs. Finalement, quatre Belles se sont endormies

dimanche avec les mêmes certitudes et les mêmes croyances initiales au sujet du déroulement de l'expérience ; elles se retrouvent à trois dans l'expérience, raisonnent de la même manière, révisent des croyances de la même manière… Mais sont-elles toujours sur une égalité épistémique ?

Soient $C_{Aur}$ la mesure rationnelle du crédit qu'Aurore accorde à ses croyances au réveil après avoir constaté que Norah dort, $C_{Est}$ la mesure rationnelle du crédit qu'Estelle accorde à ses croyances au réveil après avoir fait la même constatation. Soient les hypothèses suivantes :

A : « Aurore est une désignée »

E : « Estelle est une désignée »

J : « Je suis une désignée »

Si les Belles sont tiéristes, $C_{Aur}(J) = C_{Aur}(A) = 1/3$ et, comme seulement deux autres Belles sont réveillées et qu'Aurore n'a pas de raison de croire que l'une est désignée à un degré différent de l'autre, $C_{Aur}(E) = C_{Aur}(\neg A) / 2 = (1 – C_{Aur}(A)) / 2 = 1/3$. De même, $C_{Est}(J) = C_{Est}(E) = 1/3$ et $C_{Est}(A) = 1/3$. Aurore et Estelle croient au même degré diverses hypothèses, on ne trouvera pas de croyances liées à l'expérience sur lesquelles elles seraient en désaccord, elles peuvent dialoguer et se confirmer mutuellement leurs estimations sans problème.

Si les Belles sont demistes, $C_{Aur}(J) = C_{Aur}(A) = 1/2$ et $C_{Aur}(E) = C_{Aur}(\neg A) / 2 = (1 – C_{Aur}(A)) / 2 = 1/4$. De même, $C_{Est}(J) = C_{Est}(E) = 1/2$ et $C_{Est}(A) = 1/4$. Voilà qui est intéressant. Aurore est amenée à croire l'hypothèse (non essentiellement) indexicale J au même degré qu'Estelle, mais le pronom « Je » n'exprime pas la même personne dans la bouche d'Aurore et dans la bouche d'Estelle. Une fois que « Je » est remplacé par un nom propre non ambigu, on voit clairement que les deux amies ne partagent pas les mêmes croyances. Pourtant elles les partageaient

dimanche, et leurs cheminements mentaux jusqu'à maintenant sont identiques !

C'est la force de la variante des Quatre Belles. Le raisonnement bayésien basé sur l'inertie doxastique, fondateur de tous les courants demistes, amènerait Aurore à estimer des probabilités seulement personnelles alors qu'on s'attendrait à ce qu'elles soient interpersonnelles. Aurore doit-elle se convertir au tiérisme ou doit-elle continuer à croire en privé à des degrés qui ne sont pas ceux de ses amies ? Notons que la désambiguïsation est également et enfin contrariée, pas dans ses intuitions fondamentales mais dans ses résultats probabilistes : la variante de Titelbaum s'attaquait à des probabilités d'hypothèses temporelles que les désambiguïseurs probabilisent comme des tiéristes, mais ici A et E sont des hypothèses éternelles probabilisées selon le double demisme.

### 1.5. Des paris favorables au demisme

La variante que nous venons d'analyser est favorable au tiérisme, mais elle n'est pas déterminante. Nous avons croisé dans le dernier chapitre des discussions de chercheurs au sujet des paris hollandais. Personne ne semble proposer un système de paris efficace. Le problème est évidemment le nombre d'occasions d'offrir un pari à la Belle, qui n'est pas le même en cas de pile et en cas de face. Il apparait nécessaire qu'elle ait connaissance dimanche de la manière dont les paris sont répartis dans l'expérience, mais il ne faut pas qu'au cours de celle-ci le simple fait qu'un pari lui est proposé lui donne une information sur l'issue pile/face ou sur sa position temporelle. En outre, ne pas se restreindre aux *Dutch books* et revenir à l'idée classique qu'une disposition à parier est en relation avec l'intensité d'une croyance, peut être bénéfique.

La plus simple des solutions est aussi la plus mauvaise. Imaginons. La Belle se prépare à participer à l'expérience originale ; on l'informe alors

qu'à chacun de ses réveils elle pourra miser 10 € sur l'issue qu'elle souhaite, et retrouver sa mise plus un bénéfice de 10 € si elle devine juste. Un raisonnement naïf serait celui-ci : si la Belle mise toujours sur face, elle aura mercredi un bénéfice de 10 € si face est venu mais une perte de 20 € si pile est venu (puisqu'il y a deux occasions de parier au lieu d'une si pile), tandis que si elle mise toujours sur pile, elle aura un bénéfice de 20 € si pile, une perte de 10 € si face, ce qui montre à l'évidence qu'elle doit parier sur pile et donc, si disposition à parier et intensité d'une croyance sont liées, elle doit croire pile plus probable que face au cours de l'expérience. On voit que la conclusion d'un tel raisonnement est mal inférée et inacceptable : cette correspondance probabilité-pari manque les subtilités de la situation de l'agent rationnel. Qu'elle soit demiste ou tiériste ou autre, la Belle sait dimanche qu'il lui faudra parier sur pile pour *espérer* un gain meilleur à la fin. Au réveil dans l'expérience, certes, si elle estime à 1/2 la probabilité de face comme la probabilité de pile, son espérance de gain pour *ce* pari particulier (sur pile comme sur face) est nulle, ce qui devrait ne pas l'inciter à parier sur pile plutôt que sur face *si elle se restreint à cette seule donnée*. Mais elle sait toujours, tout autant que dimanche, que le pari lui est proposé deux fois en cas de pile ; elle sait aussi que si un pari sur pile lui apparait *aujourd'hui* comme l'action la plus rationnelle, qui lui fait espérer le meilleur bénéfice à la fin, il en sera de même lors d'un éventuel autre réveil dans l'expérience. C'est pourquoi, si elle parie sur pile, aujourd'hui et systématiquement durant toute l'expérience, son espérance de gain total, assimilable à la somme moyenne perçue par expérience (si les expériences étaient répétées), est $1/2 \times 10 \times 2 - 1/2 \times 10 = 5$ €[184]. Pour un agent rationnel, croire en face au degré 1/2 n'est pas contradictoire avec sa certitude qu'il doit parier sur pile pour espérer un bénéfice, pas plus d'ailleurs que croire en face au degré 1/3.

[184] C'est la probabilité de pile au réveil multipliée par 10 € multipliée par 2 réveils, moins la probabilité de face au réveil multipliée par 10 € (multipliée par 1 réveil).

À ce moment-là, nous croyons peut-être qu'il est vain de chercher un système de paris susceptible d'ébranler les certitudes d'un demiste ou bien d'un tiériste. Ce n'est pas vain si nous élaborons une variante qui fait conclure des probabilités plus écartées que 1/2 et 1/3, et si nous continuons à proposer des paris à chaque réveil, comme le voudrait un tiériste, mais en adaptant la mise ou la conséquence du pari en fonction du résultat du Tirage, ce que préférerait un demiste. Peut-être même en changeant la nature du pari… Examinons d'abord la variante ainsi résumée :

| Jour / Tirage | **Lundi** | **Mardi** | **Mercredi** | **Jeudi** |
|---|---|---|---|---|
| **1** | 1 | | 0 | |
| **2** | 1 | | 0 | |
| **3** | 1 | 1 | 1 | 1 |

Un tiériste (du problème original) engagé dans une telle expérience croit en l'issue 3 au degré 2/3, puisque le tableau indique deux fois plus de réveils-3 que de réveils-1 et réveils-2 réunis. Un demiste croit quant à lui en l'issue 3 au degré 1/3 seulement, puisque c'est sa croyance du dimanche et qu'il ne la modifie pas. Voici maintenant un type spécial de « pari » qui fait partie des règles très sévères connues par le sujet de l'expérience dès dimanche : à chaque réveil (il y en aura un ou quatre), et avant que le sujet ait pu glaner des informations sur le résultat du Tirage ou sur sa position temporelle, on exige de lui qu'il dise « 3 » ou « Non-3 » et pas une autre réponse. On note sa réponse. Vendredi, avant de le libérer, on triple sa fortune personnelle si le Tirage a amené 3 et s'il a répondu « 3 » à *chacun* de ses réveils, ou bien si le Tirage a amené 1 ou 2 et s'il a répondu « Non-3 » à chacun de ses réveils (le seul réveil du lundi ici). Dans le cas contraire, on divise sa fortune personnelle par trois.

Un tel vilain scénario plaira vraisemblablement à un demiste, pas à un tiériste. Et celui-ci aura beau se plaindre que les « paris » ont considérablement changé, sont injustes ou n'ont plus rien à voir avec des probabilités, il n'effacera pas (et c'est ce qui compte !) le dilemme perturbant causé au cours de l'expérience par une probabilité de 2/3 attribuée à l'issue 3. En effet, le dimanche, la Belle qui accepte de se prêter au jeu estime à seulement 1/3 la probabilité de l'issue 3, elle tient aussi pour certain qu'une fois engagée dans l'expérience il lui faudra répondre « Non-3 » quand on lui imposera le choix binaire ; elle est certaine que c'est la meilleure stratégie, la seule qui lui permet d'*espérer* être enrichie et non appauvrie à la fin, bien qu'elle ne puisse jamais être assurée de gagner. Au cours de l'expérience, elle conservera cette stratégie si elle est demiste. Si elle est tiériste, c'est plus compliqué : la première stratégie est concurrencée par une autre, par *l'autre*, son contraire. La probabilité de l'issue 3 est révisée à présent, elle est de 2/3. Si la Belle fait aujourd'hui le mauvais choix entre « 3 » et « Non-3 », elle est nécessairement ruinée, pour ainsi dire, une seule mauvaise réponse la condamnant à une forte perte ; si elle fait aujourd'hui le bon choix, elle a de bonnes chances de se retrouver riche vendredi, il suffit que sa décision soit aussi celle de toutes ses parties temporelles (réveillées) dans l'expérience, et ce sera le cas si c'est la plus rationnelle des décisions. Mais dans ce cas, c'est la réponse « 3 » qui s'impose, c'est la seule qui permet d'espérer l'enrichissement. La Belle a beau réfléchir au protocole, au fait qu'elle a quatre décisions binaires à prendre dans un monde-3, au lieu d'une dans les autres mondes, rien n'y fait : à côté de sa première stratégie qui consiste à répondre « Non-3 », il y a maintenant celle qui consiste à répondre « 3 ». Et il va sans dire que déterminer aléatoirement quoi répondre, c'est risquer encore plus la quasi-ruine. Il y a donc dilemme (au sens commun du terme, qui n'est pas le sens du logicien). Comment le surmonter ? La Belle se dira peut-être qu'elle avait une certitude dimanche : dire « Non-3 » est une meilleure action. Cette certitude n'a pas pu être brisée une fois le protocole enclenché. Il

semble judicieux de rester sur la première stratégie. Mais si elle est tiériste, elle ne pourra pas s'empêcher de croire en sa quasi-ruine au degré 2/3 une fois qu'elle aura prononcé « Non-3 », elle regrettera ce choix, c'est anormal. À moins qu'elle ne se convertisse au demisme en pensant que ce qui est une erreur dans son précédent raisonnement, c'est cette probabilité 2/3.

Croire en l'issue 3 au degré 2/3 amènerait une certitude qui contredirait celle du dimanche, ou produirait un dilemme, ou une dissonance cognitive, pour reprendre un terme plus psychologique. Le tiérisme pourrait objecter que cette conséquence fâcheuse est celle de cette variante, mais que d'autres variantes avec d'autres paris sont à l'inverse favorables au seul tiérisme. Eh bien, nous attendons qu'il découvre une telle variante, et ce ne sera pas facile. Il pourrait encore objecter que la variante ne prouve rien parce que le système de paris doit être adapté et appliqué au seul problème original de la Belle au bois dormant. Mais non : les méthodes, comptes, calculs, arguments tiéristes sont clairement contrariés par cette variante : aucun tiériste ne montrera par ses raisonnements habituels que la probabilité de l'issue 3, au réveil, est différente de 2/3.

### 1.6. La variante des Six Belles

Il faut être clair : il est très possible que la variante du dilemme tiériste que nous venons de présenter et la variante des Quatre Belles apparemment défavorable au demisme, souffrent d'une critique affutée dans un avenir proche. Mais par chance, pour tenter de dénouer le paradoxe, nous n'avons pas besoin qu'elles soient parfaites. D'ailleurs, qu'une analyse déclare vainqueurs les pro-1/2 alors qu'une autre sacre les pro-1/3 ruine l'espoir d'une infaillibilité… à moins que les deux aient raison en même temps, mais comment ? Ce qui serait une vraie curiosité, ce serait un croisement

entre une variante qui invite un agent et ses « compagnons » à harmoniser leurs croyances, et une variante de « paris-choix rationnels » qui fait détester les dilemmes intérieurs. Un peu d'astuce doit suffire. Il faut garder de chaque variante les propriétés qui les rendent si destructrices des convictions demistes ou tiéristes. Nous n'avons pas le choix, il nous faut six Belles cette fois-ci :

| Jour / Tirage | **Lundi** | **Mardi** | **Mercredi** | **Jeudi** |
|---|---|---|---|---|
| **Désignée 1** | 1 | 0 | | |
| **Désignée 2** | 0 | 1 | 0 | |
| **Désignée 3** | 0 | | 1 | 0 |
| **Désignée 4** | 0 | | | 1 |
| **Non-désignée 1** | 1 | 1 | 1 | 1 |
| **Non-désignée 2** | 1 | 1 | 1 | 1 |

Dimanche, les six Belles savent que quatre d'entre elles seront désignées par le sort pour être réveillées une seule fois durant l'expérience, les jours indiqués dans le tableau. Les deux autres seront réveillées quatre fois. De cette manière, chaque jour de l'expérience, trois Belles seront réveillées et réunies pour un entretien, quand trois autres resteront endormies. Dimanche, les six savent aussi que chaque jour on soumettra chaque réveillée au choix binaire « Désignée » / « Non-désignée ». La fortune personnelle des Belles désignées qui auront répondu « Désignée » lors de leur unique réveil, et des Belles non désignées qui auront répondu « Non-désignée » quatre fois, sera triplée vendredi ; la fortune des autres sera divisée par trois.

Nous sommes suffisamment aguerris à présent pour assurer qu'une Belle qui réussirait à faire fi de ses cinq amies et du choix binaire croirait au réveil qu'elle est une désignée au degré 2/3 si elle est une demiste (du problème original) ou au degré 1/3 si elle est tiériste. Le problème est de savoir ce qui se passe lorsqu'elle prend en compte tous les aspects de sa situation. Cette variante monstrueuse, que nous allons mettre de côté durant quelques pages, ne nous montrera pas si la ruse des compagnons l'emporte sur la ruse des paris, ou l'inverse ; en revanche, elle peut nous faire osciller entre les probabilités 1/3 et 2/3, elle peut faire découvrir aux partisans d'une solution le point de vue adverse. Car la force des paradoxes enferme dans la partialité des gens d'habitude ouverts d'esprit et capables de s'adapter, elle les empêche de basculer dans l'étrangeté de l'autre. Les chercheurs témoignent par leur souci de synthétiser des vues adverses que ce n'est pas la beauté du paradoxe de la Belle au bois dormant qu'ils ont découverte en alternant les perspectives, mais la solution elle-même, encore très peu distincte. Nous pensons en particulier aux désambiguïseurs qui ont complètement cessé de voir un tiérisme retenu par la probabilité ontique et un (double) demisme retenu par la probabilité épistémique. Un même genre de probabilité mène à tous les points de vue, et peut-être à la solution.

## 2. Le raisonnement anthropique revisité

Nous avons souvent évoqué le double demiste Nick Bostrom en même temps que les désambiguïseurs : des traits communs sont décelables, un appui sur une même intuition peut-être. Bostrom a décrit une variante de l'expérience originale qui dure un million de jours : en cas de pile la Belle est réveillée un million de fois au lieu de deux fois[185]. Ce scénario extrême,

[185] C'est l'*Extreme Sleeping Beauty* de Bostrom (2007). Précisément, la Belle est réveillée un million et une fois.

selon son auteur, rendrait encore plus intuitives les conclusions doubles demistes. On comprend ce qui gêne un habitué du raisonnement anthropique dans la façon de penser tiériste. Supposons que nous croyions à un certain degré non extrême que notre conscience apparait à notre naissance physique pour être anéantie à notre mort ; supposons que nous croyions à un degré là encore non extrême cette autre hypothèse selon laquelle notre conscience dure éternellement, s'étendant dans des vies physiques innombrables, mais oubliant à chaque renaissance la vie antérieure. Arrive un tiériste audacieux qui déclare que nous n'avons pas pris en compte une noble information : *nous sommes des observateurs dans un état conscient actuellement*, ce qui rend impossible la première hypothèse et, en l'absence d'une troisième théorie suffisamment compétitive, augmente jusqu'à 1 la probabilité de la seconde hypothèse, qui invite à probabiliser une infinité de centres. Pourtant les tiéristes ne sont pas tous prêts à accepter ce raisonnement. Ce qui manque dans leur argumentation est un éclaircissement de ce qui distingue l'aventure de la Belle et la méditation d'une âme qui émet l'hypothèse de sa persistance au-delà d'une vie limitée par une naissance et une mort.

Les fréquentistes répondront peut-être qu'il n'y a rien de comparable entre un problème de type Belle au bois dormant, concret et précis, qui décrit le dispositif aléatoire dans le moindre détail et dont l'expérience est répétable, et un problème qui décrit une situation unique, où sont mêlées la métaphysique et la spiritualité, fait pour la croyance dans tous les sens du terme. Pourtant, le raisonnement anthropique réclame parfois la reconstitution d'un dispositif aléatoire subtil et la lecture de probabilités ontiques pour justifier des croyances. Nous traitons deux exemples.

### 2.1. Londres et Petit-Bled

Londres et Petit-Bled (*London and Little Puddle*) est une expérience de pensée inventée par John Leslie, qui la place comme intermédiaire entre un problème probabiliste indéniablement analogue à un problème d'urne et un problème beaucoup plus délicat comme l'argument de l'Apocalypse (qui doit aussi avoir son analogue, selon l'auteur) :

> Regardez encore votre soi amnésique, quand vous essayez de dire où vous avez le plus de chances d'être, à Londres ou bien à Petit-Bled. Pour vous guider, vous n'avez rien d'autre que les chiffres de la population. Pour simplifier les choses, supposez que […] tous les êtres humains doivent se trouver à Petit-Bled ou bien à Londres, et vous le savez. Vous savez aussi que les populations sont cinquante habitants et dix millions, respectivement. Un modèle approprié de la situation est une urne contenant cinquante boules marquées « Petit-Bled » et dix millions marquées « Londres ».[186]

Leslie raisonne parfois en bayésien. Dans le vocabulaire qu'il emploie ici, dans le raisonnement qu'il prépare, on comprend qu'il a une plus immédiate lecture ontologique de la probabilité. Le principe épistémologique d'indifférence devient superficiel, les possibilités « Londres » et « Petit-Bled » n'ont pas le temps logique d'être équiprobables, la taille et le rapport des populations accaparent toute l'attention, un nombre mesuré dans le microcosme Londres-Petit-Bled ou le résultat d'un calcul employant de tels nombres est prêt à devenir la première estimation probabiliste. Ce microcosme n'est même plus seulement analogue à une urne, il est une urne découverte derrière la situation d'un sujet amnésique qui essaie de trouver sa place dans le monde : le sujet est invité à aligner sa croyance en la localité « Petit-Bled » sur les chances objectives, pour une « main aveugle » qui se saisit d'un habitant quelconque du microcosme, de saisir un habitant de Petit-Bled. Et les chances sont minces étant donné l'immense population de Londres.

[186] Leslie (1996), p. 209.

Quand l'amnésique croit qu'il se trouve probablement à Londres, c'est qu'il a distingué un dispositif aléatoire, l'agencement de ses éléments, les lois qui le régissent : il se perçoit en habitant isolé parmi dix millions et cinquante habitants qui ont chacun une existence propre, et il ne voit aucune raison pour que les habitants de Petit-Bled aient plus de chances que les autres de souffrir d'amnésie et de s'interroger sur leur localisation. Pourtant il n'est pas clair que toute subjectivité croirait au même degré qu'elle est à Londres, car la connaissance du dispositif aléatoire n'est pas parfaite. Que se passerait-il si, comme le suggère un peu plus loin Leslie, notre soi amnésique apprenait que Petit-Bled existe réellement mais que Londres *peut* ne pas exister ? Nous augmenterions la probabilité d'être à Petit-Bled, à tel point que si nous nous rendions compte soudain que nous habitons à Petit-Bled, nous serions presque certains que Londres n'a aucune existence.

Leslie insiste sur le fait que lorsque Londres et Petit-Bled existent *tous les deux* dans notre espace-temps et qu'on le sait, il est normal de croire très fortement que l'on habite à Londres. C'est lorsque les populations se déversent dans l'espace logique, lorsqu'une population *peut* être de telle taille ou de telle autre, que tout se complique[187].

### 2.2. Le Pile ou face divin

Le Pile ou face divin apparait chez Leslie, est repris et adapté par Bostrom dans sa dissertation doctorale. Nous nous concentrons uniquement sur la version de Leslie, la plus ancienne :

> Supposez que tous les êtres humains devaient exister au même moment. Supposez que vous ne savez pas du tout s'il existe d'autres humains à part vous-même. Vous savez seulement que Dieu a décidé de lancer une pièce de monnaie *juste une fois*, et que si elle est tombée sur face il a créé quatre-vingt-dix millions

[187] *Ibid.*, p. 228.

d'humains, mais que si elle est tombée sur pile il a créé un unique être humain. Penseriez-vous alors que les chances que la pièce de Dieu soit tombée sur face sont de quatre-vingt-dix millions contre une ?

En guise de variante, supposez que l'issue face aurait conduit à quatre-vingt-dix millions d'individus, l'un d'eux portant le nom « Dr Lenoir », alors que l'issue pile aurait conduit à une seule personne, « Dr Leblanc ». Ayant oublié votre nom, mais sachant que vous avez été créé à la suite du lancer, penseriez-vous qu'il est tout aussi judicieux de parier que vous êtes Dr Lenoir et de parier que vous êtes Dr Leblanc ?[188]

Encore plus abstraite que Londres et Petit-Bled, cette expérience de pensée est aussi plus clivante. Leslie, qui répondrait non aux deux questions, essaie de montrer qu'il n'y a pas plus de chances d'exister au sein d'une population humaine très large que de chances d'exister dans une population réduite à l'extrême, autrement dit, avec les mots de Bostrom, qu'il faut rejeter l'hypothèse anthropique SIA[189] (et accepter l'argument de l'Apocalypse). Il n'est pas sûr que Leslie puisse convaincre tout le monde. Par exemple, il y a une ruse derrière la variante Lenoir-Leblanc : si j'entends deux voix, une qui me dit qu'un pari sur Lenoir vaut le même pari sur Leblanc, l'autre qui me dit qu'il vaut largement mieux miser sur Leblanc, je suis tenté de miser sur Leblanc même si je ne sais pas si c'est *vraiment* plus judicieux.

Néanmoins il faut reconnaitre que le Pile ou face divin a des chances de convaincre certains chercheurs auparavant tentés de raisonner sur le *Doomsday Argument* de la même façon que sur Londres et Petit-Bled. On a envie d'accepter au moins la conclusion de Leslie ici, quoiqu'avec réticence. C'est intéressant de savoir pourquoi, car la ressemblance avec la Belle au bois dormant se fait plus précise.

---

[188] *Ibid.*, p. 227.

[189] L'hypothèse *Self-Indication Assumption* de Bostrom est formulée dans notre chapitre 2, mais la formulation approximative de Leslie nous en rappelle l'idée.

Pour le philosophe canadien, le Pile ou face divin perd quelque chose de Londres et Petit-Bled pour se rapprocher de l'argument de l'Apocalypse : c'est l'effectivité des populations mises en concurrence, lesquelles sont maintenant seulement possibles. Ou bien ce monde compte un individu, ou bien il en compte des millions. Par la pensée, le sujet ne se déplace plus seulement dans le monde (d'un individu à l'autre), mais aussi dans les mondes (l'espace logique), qui auraient en outre une sorte de primauté : les probabilités s'y distribueraient avant de se distribuer sur les centres. Nous reconnaissons un peu l'argumentation utilisée par le philosophe pour la Belle au bois dormant, qu'il analyse en demiste. Cependant, il est douteux que la différence effectivité/possibilité ou existence/inexistence explique la distribution des probabilités. En effet, si l'on se tourne justement vers la Belle au bois dormant, on constate que les tiéristes sont majoritaires alors que vraisemblablement la plupart d'entre eux n'adhèrent pas au réalisme modal qui confère l'existence aux mondes-pile et aux mondes-face. Et David Lewis, qui adhère au réalisme modal comme personne, est demiste ! On s'attendrait à la situation inverse. Leslie serait donc en train de déjouer les intuitions de dizaines de chercheurs aguerris. Peut-être… mais il y a une autre solution.

Le Pile ou face divin acquiert une différence à la fois par rapport à l'Apocalypse et par rapport au Londres : le jet de la pièce de monnaie, extérieur au monde créé, est au cœur du dispositif aléatoire. La connaissance de ce lancer divin équitable par le sujet qui ignore s'il est seul au monde l'incite à penser qu'il est privilégié ou que son existence dans ce monde était une nécessité ou même qu'il a préexisté avec Dieu d'une manière ou d'une autre. Ainsi il vit dans « le temps de Dieu » et croit comme Dieu (et comme Leslie) en l'équiprobabilité de pile et de face. Pourtant, le dispositif aléatoire divin pouvait être tout autre : le sujet aurait pu ne jamais exister ailleurs que dans les plans de son créateur, et le fait qu'il soit aujourd'hui une créature devrait alors lui faire croire très

fortement en face, issue qui amène des millions d'autres créatures. Supposons malgré tout que l'existence du sujet en ce monde créé était nécessaire, c'est-à-dire voulue par Dieu avant même le tirage à pile ou face ; nous le supposons parce que nous nous approchons ainsi de la situation de la Belle au bois dormant qui a une existence avant (et après) l'expérience et une existence dans le cours de l'expérience, vie particulière rythmée par l'amnésie. Une analyse subtile et difficile commence maintenant. Le sujet habite peut-être ce monde en compagnie de millions d'autres créatures qui ont les mêmes facultés mentales que lui, et il ne peut être un élu de Dieu au milieu d'elles. Ce qu'il faut plutôt penser est que toutes ces créatures ont une existence nécessaire, elles seraient venues à l'existence quelle que soit l'issue du tirage divin. Mais cela change tout puisque dans un monde-pile il n'y a qu'une créature : cela signifie qu'il n'y a jamais eu qu'une seule créature même si *paradoxalement* il y en a des millions dans le monde-face, autrement dit Dieu n'a pas simplement créé un ou des individus, mais il a créé un individu occupant dans le monde-face de multiples places qui, elles, n'ont pas l'existence nécessaire même si elles se rassemblent toutes dans un seul être nécessaire. Le sujet a deux plans d'existence : celui où il est unique, indivis et dans « le temps de Dieu » et celui où il est une division, une partie parmi les parties au nombre de 1 dans le monde-pile ou de $9.10^7$ dans le monde-face.

Si l'on soutient cette analyse (et ce ne sera pas sans mal), Leslie a raison et tort en même temps. D'où notre acceptation et notre réticence devant ses conclusions. Le sujet, créature ambigüe ou paradoxale de Dieu, peut avoir deux lectures ontologiques des probabilités : il peut exister comme la créature indivise présente une fois dans le monde possible où pile est venu, une fois dans le monde possible où face est venu, et alors lire la proportion 1/2 (une créature-face sur deux créatures indivises possibles), puis croire en face au degré aligné 1/2 ; il peut aussi exister sur le plan de la multiplicité comme un être différent des autres habitants (éventuels) du

monde, lire la proportion $9.10^7 / (9.10^7 + 1)$ et donc avoir une quasi-certitude que face est venu. Il faut considérer que ce dernier plan d'existence, qui ici déséquilibre les populations des mondes possibles, tord le processus aléatoire *a priori* équitable pour les mondes, jusqu'à déplacer l'équité sur les mondes *centrés*, jusqu'à faire de l'habitant du monde-pile une identité tout aussi probable, pour notre sujet qui s'interroge, qu'un des nombreux habitants du monde-face.

Voilà une histoire incroyable, dira-t-on. En énonçant le Pile ou face divin, Leslie ne voyait pas ces subtilités métaphysiques, ces êtres ambigus, à la fois nécessaires et contingents ; il simplifiait la création du monde par un dieu, il en faisait un processus aléatoire détaillé, il voulait éliminer tout mystère. Eh bien, manifestement, Leslie n'a pas été assez précis, *puisque le Pile ou face divin est un problème dont la solution n'est pas claire*. Il n'a pas réduit la création du dieu à un processus détaillé. Un des habitants créés croit avec raison en face au degré 1/2 ? Peut-être bien, mais cela implique son existence nécessaire, détail tu par Leslie. Nous ne divaguons pas, nous réagissons du mieux que nous pouvons à un énoncé laissé dans l'imprécision. Pourtant, dira-t-on encore, cette réaction mène à tout un discours sur la nécessité, les « plans d'existence », le « temps de Dieu »… Est-ce raisonnable de vouloir apporter une telle solution à un problème à énoncé ambigu, au lieu de simplement relever l'ambiguïté ? Mais tout ce que nous disons, c'est que Leslie n'a pas été assez précis ; nous ne disons pas qu'il aurait pu l'être, nous ne croyons pas qu'il aurait vraiment pu réduire à un dispositif à la simplicité enfantine une telle création d'une population aléatoire. Les plans d'existence ne sont pas la conséquence d'une ambiguïté de l'énoncé. Attention : nous parlons bien du monde fictif du Pile ou face divin, nous ne nous sommes jamais lancés dans un discours passionné sur le monde réel et notre présence en son sein.

## 3. Dénouement

On peut encore admettre que le monde du Pile ou face divin est régi par des lois ontologiques singulières. Mais trouver des lois similaires dans la Belle au bois dormant est *a priori* douteux. S'il y a un dieu, il ne joue aucun rôle ici ; aucun agent rationnel n'est « créé » ou anéanti ; on imagine même l'expérience de la Belle reproduite dans un laboratoire. Et pourtant, la Belle est assez manifestement un être paradoxal : elle est psychologiquement continue par nature mais l'amnésie engendre une discontinuité entre la Belle du lundi et la Belle du mardi, laquelle vit son réveil comme si c'était le premier, « recommence » le précédent réveil sans en avoir conscience. La Belle n'a pas été découpée en plusieurs individus, mais ses parties temporelles n'ont jamais été aussi séparées et distinguées. D'accord, dira-t-on, mais il y a un grand pas à franchir avant de dire que l'expérience est comme une dénaturation de l'être de l'agent, sa recréation sur deux plans d'existence ! Bien plutôt, le parangon de rationalité observe tout, déduit tout, sait tout ce qu'il est possible d'atteindre dans sa situation, il n'hésite pas entre une prise en compte de ses parties temporelles ou leur dissolution dans son être continu. Peut-être, peut-être pas…

### 3.1. Retour sur la variante des Six Belles

Nous pensons que la variante des six Belles est révélatrice. Suivons Aurore, décidée à participer à l'expérience avec cinq amies. Avant de s'endormir dimanche, elle croit au degré 2/3 qu'elle sera désignée par le sort pour n'être réveillée qu'une seule fois, et elle sait que la stratégie pour avoir les meilleures chances de s'enrichir, quand on lui demandera de prononcer « Désignée » ou « Non-désignée », est de dire « Désignée ». Supposons qu'elle n'altère ni cette croyance ni cette connaissance au réveil, et plus que ça, faisons l'effort de nous mettre à sa place et de faire comme elle, quel que soit notre positionnement dans la controverse *Sleeping*

*Beauty*. Les expérimentateurs réunissent les trois Belles réveillées, lesquelles déduisent quelles sont les trois Belles dont le sommeil n'a pas été interrompu aujourd'hui et qui sont nécessairement des désignées. Les réveillées savent qu'une seule réveillée est une désignée. Imaginons la conversation qui s'ensuit et l'incompréhension causée par des croyances mal mesurées. Aurore trouve intenable la probabilité 2/3 qu'elle associait à « Je suis une désignée », elle ne peut pas se résoudre à croire si faiblement (au degré 1/6) que par exemple son amie Estelle, son égale face au protocole, est désignée. Si les Belles participaient à de multiples expériences, Aurore serait la désignée dans une sur trois rencontres avec ses deux amies. Elle doit croire au degré 1/3 qu'elle est désignée, au degré 1/3 qu'Estelle est désignée, au degré 1/3 que la troisième réveillée est désignée ; toutes les réveillées harmonisent ainsi leurs croyances. Nous aussi, basculons dans ce tiérisme et ressentons intimement le charme qui nous y retient.

Peu de temps après, les expérimentateurs demandent aux trois Belles de se séparer et de retourner dans leurs chambres. Puis ils proposent à Aurore le choix binaire : « Désignée » ou « Non-désignée » doit maintenant sortir de sa bouche. La demoiselle est dans l'embarras, car une seconde stratégie lui est apparue : dire « Non-désignée ». Elle croit en effet qu'elle n'est pas désignée au degré 2/3 ; c'est suffisant quand on sait que donner une réponse conforme à son statut est se donner de bonnes chances de s'enrichir, alors que se tromper conduit nécessairement à une énorme perte correspondant à deux tiers de sa fortune personnelle. Aurore se ressaisit : elle se dit que si toutes les Belles optaient pour « Non-désignée », il n'y aurait que deux riches vendredi, une fois l'expérience achevée ; il vaudrait donc mieux revenir à la première stratégie. En réfléchissant aux deux stratégies opposées, il doit lui apparaitre, et à nous aussi, que le degré 2/3 ne peut pas être accordé à « Je ne suis pas désignée », mais bien plutôt, comme au début, à la croyance contraire « Je suis une désignée » ; après

tout, quatre Belles sur six sont désignées, Aurore serait désignée dans deux expériences sur trois si les expériences étaient répétées. Elle répond donc « Désignée » conformément à sa première stratégie, et elle le fait sans regret et sans crainte, car elle a rebasculé vers ses croyances initiales.

Certes, l'Aurore dont nous racontons la mésaventure n'est pas un parangon de rationalité ; celui-ci réfléchirait bien plus vite, anticiperait les événements, et pourtant nous ne croyons pas qu'il trouverait la sortie d'une oscillation entre deux probabilités. Ou alors il se stabiliserait sur une des deux, mais pas parce qu'elle est absolument meilleure que l'autre.

Il est temps d'essayer de mieux comprendre le curieux mécanisme du « basculement » de 1/3 vers 2/3 et inversement. Il ne s'agit pas d'une révision d'une croyance initiale après perte ou gain d'information, presque tout ce qui arrive à Aurore était par elle attendu et les certitudes qu'elle peut acquérir ne sont pas suffisantes pour la maintenir sur 1/3 ou sur 2/3 ; quoique conditionaliser soit parfois tentant selon les analyses et les points de vue, quelque chose comme une « déconditionalisation », un retour à la croyance *a priori*, semble toujours possible. En même temps, les deux probabilités épistémiques concurrentes ne sont pas du tout des caprices radicalement subjectivistes et la raison, la réflexion est sollicitée en permanence dans ces basculements. La raison se porte tout à tour sur telle ou telle proportions, observons bien : une Belle sur trois Belles aujourd'hui réveillées est une désignée, quatre Belles sur les six (réveillées ou non) sont des désignées… Il n'y a pas d'autre solution : le basculement est détachement d'une probabilité ontique ancienne et, simultanément, réalignement sur une probabilité ontique nouvelle. C'est un déplacement sur le support ontique apparemment ordonné par le fait que la Belle « ne se considère plus comme » (pour le dire ainsi provisoirement) une « réveillée ou non » dans son être continu que l'expérience et l'amnésie ne sauraient diviser, mais comme une « réveillée aujourd'hui », en tant que partie temporelle qui se localise autant qu'elle peut. Ou c'est le retour en arrière,

ordonné par le fait inverse. Ces probabilités ontiques ne sont peut-être pas simplement les proportions ou fréquences ci-dessus mentionnées, il nous reste quelques subtilités à découvrir.

### 3.2. Le déphasage de la Belle au bois dormant originale

La variante des Six Belles force le basculement en plaçant à tour de rôle le demisme et le tiérisme en grande difficulté. Elle éclaire ainsi le problème original, où ce phénomène n'est pas apparent. La solution que nous proposons ressemble à la désambiguïsation parce que la Belle doit répondre 1/2 *ou* 1/3 à la question principale, 1/2 à la question subsidiaire. Toutefois elle s'écarte de la désambiguïsation sur bien des points importants, notamment elle ne croit pas qu'une ambiguïté doit être cherchée dans l'énoncé et ne conteste pas le principe d'équiprobabilité des propositions équivalentes étendu aux propositions auto-localisantes.

Commençons par une définition. Nous pensons que nous sommes une succession de tranches temporelles, plus précisément nous sommes un seul individu subsistant mais aussi une de ses parties au temps « maintenant », et généralement nous passons du tout à la partie sans nous en rendre compte, les deux étant « en phase » épistémiquement et doxastiquement. Pour qu'un agent rationnel ait possibilité de basculer, il faut au préalable produire un certain bouleversement ontique et épistémique : un *déphasage* de l'agent, c'est-à-dire, comme l'a entrevu Bostrom selon nous, une discordance entre son tout et ses parties, entre l'être continu et ses centres. Un déphasage, c'est à l'origine un décalage de deux phénomènes alternatifs de même fréquence : si les deux sont en phase, ils seront déphasés dès que la fréquence de l'un va légèrement varier, au moins momentanément. Au sens figuré, le déphasage d'un sujet est une perte de contact avec la réalité, par exemple à la suite d'un voyage et d'un décalage horaire. Le déphasage dont nous parlons à présent est une perte de contact de l'agent avec un

aspect de sa propre réalité, être continu ou bien partie temporelle, au profit de l'autre aspect ; il se produit lorsque des données extraites du monde, notamment les proportions et les fréquences relatives susceptibles de fonder des crédits, ne sont plus les mêmes pour l'agent dans sa dimension totale et pour l'agent dans sa dimension partielle.

Le dimanche soir, la Belle est dans son état normal, pour ainsi dire. Sa situation dans le monde n'est pas pertinente pour ses croyances éternelles et notamment pour l'issue pile/face. Elle peut ne pas connaitre exactement l'heure qu'il est, mais cela n'a aucune importance parce qu'elle n'envisage pas des mondes-face abritant des centres possibles en nombre différent du nombre des possibles dans les mondes-pile. Elle sait qu'au cours de l'expérience, ce ne sera pas la même chose ; c'est là encore sans conséquence pour l'état présent de ses croyances. Lundi, on la réveille. Elle n'a pas encore pris la drogue mais n'en sait rien, elle n'est pas amnésique mais l'effet sur les croyances temporelles est bien là : elle ignore si l'on est lundi ou mardi. Et surtout, si face est venu, se dit-elle, on est lundi, tandis que si pile est venu, on peut être lundi ou mardi. Le déséquilibre est patent, le déphasage est inévitable. La raison de l'agent va reposer sur le tout ou bien sur la partie et va être retenue à ce seul niveau ontique à cause des contradictions que le passage d'un niveau à l'autre engendrerait. Il faut donc considérer (c'est peut-être la plus grande difficulté) que la Belle qui répond 1/2 à la question principale est un agent parfaitement cohérent mais qui n'*est* pas la Belle qui répond 1/3 avec une cohérence égale. Elle est peut-être en manque d'une part d'être, elle ressent le déphasage, ce qui d'ailleurs peut lui permettre de basculer d'un niveau à l'autre et d'une croyance à l'autre, mais le déphasage est comme une fission inachevée (et inachevable) de deux identités, la Belle demiste en tant que continuité et la Belle tiériste en tant que tranche temporelle. Les raisonnements des deux identités diffèrent énormément.

La Belle tiériste se réveille comme le décrit Horgan dans un environnement centré, se localisant dans les mondes et dans le monde. Les indexicaux essentiels intègrent harmonieusement ses raisonnements. Cette attention portée sur des parties temporelles nécessairement renouvelées par le passage du temps rend caduque la probabilité de face 1/2 du dimanche. La Belle sait qu'elle est réveillée, comme le critique Pust le signale, mais elle a en vue non seulement toutes les Belles possibles, mais toutes les Belles d'aujourd'hui, un aujourd'hui qui est lundi ou mardi. Elle compte donc quatre « Belles » dont trois seulement sont possibles puisque réveillées, autrement dit elle envisage trois compartiments :

– < < Belle, lundi >, face > (possible : premier compartiment)

– < < Belle, mardi >, face > (impossible)

– < < Belle, lundi >, pile > (possible : deuxième compartiment)

– < < Belle, mardi >, pile > (possible : troisième compartiment)

Les trois possibles sont exclusifs et conjointement exhaustifs, l'équiprobabilité est de mise ; pourtant il ne s'agit pas d'une équiprobabilité épistémique d'indifférence car la Belle est dirigée ontologiquement, par ce qu'elle est et ce qu'elle compte. Elle doit repenser à la variante des Quatre Belles si elle doute de la probabilité 1/3 du premier compartiment : elle doit ajuster ses croyances comme si elle était en présence de deux compagnons réveillés, dans le même état épistémique qu'elle. Assurément, lors de rencontres renouvelées avec ces deux compagnons au cours d'expériences similaires, elle serait une fois sur trois la Belle-face. La croyance en face est alignée sur cette probabilité ontique par une inférence directe. Lorsque les expérimentateurs apprennent à la Belle qu'on est lundi, celle-ci peut conditionaliser sur cette information et retrouver la probabilité 1/2 du dimanche. Le déphasage est terminé à ce moment.

La Belle demiste (en fait, double demiste) se réveille en croyant en face au degré 1/2 parce que depuis dimanche elle continue à se reposer sur l'équiprobabilité des mondes-pile et des mondes-face, ontologiquement fondée. En tant qu'être continu, elle se décrit toujours, elle et ses contreparties, en s'associant à des événements datés, dont la probabilité ne souffre pas du simple passage du temps (la Belle-pile, la Belle des expériences-face…). Si un réveil dans l'expérience avait été un événement impossible dans certains mondes, la probabilité de face aurait pu changer, mais ce n'est pas le cas. Une information telle que « La pièce tombe sur face et on est lundi, ou la pièce tombe sur pile et on est lundi ou mardi » n'a aucune pertinence parce qu'une précision sur les positions temporelles possibles dans tel ou tel monde ne saurait signifier l'élimination d'un monde. Plus que cela, la Belle fuit les indexicaux essentiels, pas parce qu'elle est limitée (et manque des données « spécifiques », comme dirait Horgan) mais parce que c'est rationnel, parce que ses associations d'idées réclament la cohérence : elle ne produit aucun non-sens de type « Combien y a-t-il d'expériences dans lesquelles aujourd'hui est mardi ? » et préférera « Combien y a-t-il d'expériences dans lesquelles on me réveille mardi ? »

Lorsque les expérimentateurs lui apprennent qu'on est lundi, la Belle devient capable de se situer dans un jour particulier de l'expérience. L'information « Aujourd'hui est lundi » est particulière : bien qu'essentiellement indexicale, elle ruine l'essentialité de l'indexical puisque, une fois que l'agent l'a assimilée, il peut remplacer « lundi » par « aujourd'hui » dans tous ses raisonnements. Une Belle demiste entend davantage « Lundi est aujourd'hui » quand les expérimentateurs font l'annonce. Mais il faut reconnaitre qu'il y a une difficulté. Première option : l'information ne nuit pas à la cohérence des associations d'idées et ne fait que donner à la Belle la possibilité d'envisager les événements spécifiques du lundi ; ainsi elle raisonne comme dans le paragraphe précédent mais avec des lundis-pile et des lundis-face plutôt qu'avec des

expériences-pile et des expériences-face ; puisqu'un réveil le lundi est toujours prévu par le protocole, elle aligne sa croyance en face sur l'évidente fréquence 1/2 des lundis-face parmi les lundis des expériences répétées. Seconde option : l'information nuit à la cohérence parce que la Belle ressent sa pertinence (elle aurait pu apprendre qu'on est mardi, ce qui aurait annulé la probabilité de face) et a envie de conditionaliser sur elle ; elle ne peut conditionaliser que si d'abord elle bascule dans un système de raisonnements acceptant les propositions essentiellement indexicales, donc si l'être continu cède la place à la partie temporelle ; le basculement est ainsi une lecture ontologique des probabilités recentrée sur des populations et des proportions de parties et contreparties temporelles ; après alignement, les résultats épistémiques tiéristes s'imposent à la Belle ex-demiste et c'est 1/3 qui est le crédit *a priori* accordé à face, avant conditionalisation ; *a posteriori*, la probabilité de face est à nouveau 1/2. Quel que soit le cheminement mental de la Belle, la réponse à la question subsidiaire est toujours la même. Bien sûr, le déphasage disparait au moment où l'agent se localise dans le temps aussi bien que les expérimentateurs.

Le simple demisme ne voit pas la composition basculement plus conditionalisation et aboutit alors au crédit 2/3 de face (sachant lundi), crédit sans aucun support ontique. David Lewis et ses successeurs veulent seulement conditionaliser sur l'annonce des expérimentateurs parce que le basculement est chez ces bayésiens hors d'atteinte, il nécessiterait de s'interroger sur l'être de l'agent raisonnant et probabilisant. Il ne suffit pas de distinguer en théorie le *continuant* et l'*individual-at-time*, il faut surtout comprendre que l'agent déphasé qui se range lui-même dans des populations de *continuants* ayant certaines propriétés ne peut pas se ranger soudain dans des populations d'*individuals-at-times* sans bouleverser son espace de probabilité et même sa façon de raisonner, qui est toujours la « bonne » façon, propre à son être. Et nous n'acceptons ce bouleversement

que si nous considérons les agents avant et après basculement comme distincts, sinon il ne faudrait y voir que contradiction et absurdité. Pas distincts physiquement, donc toujours identifiables dans notre espace-temps comme un seul agent. C'est la difficulté.

### 3.3. La mise à mort de la scission ontique-épistémique

Une telle résolution du paradoxe apportera son lot d'objections. Concentrons-nous uniquement sur la principale, qui n'entre pas dans le détail et les considérations métaphysiques. Cette « solution », dira-t-on, est semblable à la désambiguïsation dans le sens où la synthèse promise est une tentative originale pour faire cohabiter les points de vue contraires, c'est donc le paradoxe mis sur un piédestal, une excuse ou une justification de la contradiction et non l'élimination d'un raisonnement contradictoire par la découverte d'une erreur.

Cette solution n'est ni la répétition du paradoxe, ni la répétition des arguments premiers des chercheurs. C'est tout l'opposé. Une partie de la force de la Belle au bois dormant tient dans l'écart peu banal entre deux réponses apportées à la même demande d'une estimation probabiliste : 1/2 et 1/3. 1/2 est la réponse première d'un bayésien, la réponse de qui cherche à conclure d'un argument son intuition de ce que devrait être le *degré rationnel de croyance* en l'issue face ; 1/3 est la réponse première d'un fréquentiste, la réponse de qui cherche à conclure d'un argument son intuition de ce que devrait être la *fréquence relative* de l'issue face. Dissoudre le paradoxe, c'est d'abord craindre cette scission, puis s'inscrire dans la droite ligne des recherches jusqu'ici menées pour la réparer. Notre solution aurait répété le paradoxe si elle avait consisté à dire : 1/2 est une probabilité épistémique, 1/3 est une probabilité ontique, d'un autre ordre ; nous nous inquiétons de leur écart à tort, la rationalité doit pouvoir s'en satisfaire. Or, l'article d'Adam Elga date de 2000 et depuis une centaine de

publications se sont inquiétées. C'est que le paradoxe n'est pas confusion de deux concepts de probabilité, de deux types de discours scientifiques. Cet écart est anormal. Les tiéristes et les demistes cherchent en priorité à réduire l'écart de deux *quantités*, la façon la plus simple étant de trouver la « bonne » probabilité parmi les deux prétendantes. Notre priorité est la réparation de l'écart *qualitatif*, de la scission ontique-épistémique. La solution proposée, qui s'inspire de plusieurs années de débats, va très loin dans cette réparation : nous disons que 1/2 et 1/3 sont toutes deux des probabilités épistémiques alignées sur des ontiques ; nous disons que 1/2 est la probabilité estimée par l'agent en tant que continuité, et 1/3 la probabilité estimée par l'agent en tant que partie temporelle. Rien ne saurait rendre 1/2 plus objective ou plus subjective que 1/3, les deux mesures sont à égalité qualitative. La scission est morte pour qui distingue encore deux interprétations de la probabilité, agonisante pour les affamés qui réclament davantage, à savoir la fin du Paradoxe suprême de la double interprétation.

Les désambiguïseurs ont pressenti qu'il y avait *quelque chose comme* deux bonnes probabilités de face estimées par un agent rationnel. Comme nous, ils pensent qu'il n'y a pas deux estimations valables en même temps pour le même événement (ou la même proposition) et le même individu, et que ce serait justement affirmer une telle chose qui reconduirait le paradoxe. Alors ils choisissent de distinguer dans l'événement deux événements, deux « faces » ; nous avons vu les détails dans le dernier chapitre. Cela trahit une certaine radicalité gênante, surtout chez les désambiguïseurs fréquentistes. Nous pensons que la meilleure solution consiste à dédoubler l'agent et non l'événement, aussi incroyable que cela puisse paraitre à première vue. La scission ontique-épistémique n'en est que plus facilement réparable.

# Conclusion

Les nouvelles théories de dynamique doxastique soulignant le rôle du temps ont fleuri ces dernières années. Chez Horgan, Titelbaum, Meacham ou encore Bradley… C'est propre, copieux, excitant, mais c'est aussi un peu gênant, parce que les uns ne s'entendent pas avec les autres. Ils sont trop pressés. Les problèmes d'auto-localisation sont exigeants.

Quand on affronte un paradoxe logique, hésiter et osciller entre deux propositions de solution opposées montre effectivement la beauté du problème, laquelle n'est donc pas appréciable par tout le monde. En général, cela manifeste aussi sa puissance, quand il semble tordre notre raison et fait échouer nos arguments. Mais nous avons essayé de montrer que la Belle au bois dormant fait partie de ces rares paradoxes qui trouvent leur solution dans l'acceptation totale des deux pôles de l'oscillation. C'est justement là que se dénoue l'intrication entre un discours poussant à la rupture, et non à la surveillance mutuelle voire à la conciliation, les probabilités ontique et épistémique, et un discours sur la probabilisation de nouveaux objets dont le temps est un paramètre essentiel, une intrication qui remonte peut-être à D'Alembert. L'étrangeté séduisante de la plus célèbre des énigmes d'auto-localisation incitait à croire que la dissociation complète de deux genres de probabilité était nécessaire à l'épistémologie ; les philosophes ont plus ou moins activement réagi contre cette idée, comme si elle était contre-nature, mais la tentation de la rupture aurait

ressurgi tôt ou tard, dans un des nombreux puzzles que les chercheurs continuent à inventer pour redonner de la vivacité au paradoxe. Notre solution est radicale : nous demandons de reconnaitre qu'un agent rationnel qui, en se localisant dans le monde, est soudain confronté à des écarts de probabilités et a tendance à préférer une probabilité à une autre, n'est jamais confronté à une scission ontique-épistémique qui fâche notamment les fréquentistes et les bayésiens, mais à un déphasage dans sa propre réalité paradoxale de succession *unique* de *plusieurs* tranches temporelles, un phénomène qui maintient en très bons termes les deux interprétations de la probabilité, au point que nous ne puissions plus exclure un pari sur un concept unique de probabilité, le pari de Poincaré en somme. Notre réflexion sur le déphasage appartient au seul discours sur l'auto-localisation, elle lave ce discours qui depuis longtemps entrainait avec lui un discours sur une *rivalité* ontique-épistémique.

Une approche combinée de l'interprétation double et de l'auto-localisation risquait de dérouter. Il n'y a pourtant pas d'autre moyen de résoudre un problème de *self-location* qui interroge en même temps l'essence de la probabilité. S'efforcer de penser la Belle au bois dormant comme le nœud absolu de théories qui ne se sont pas développées indépendamment, c'est regarder le paradoxe en entier, tel qu'il est : un enlisement philosophique préparé depuis des décennies. Mais justement, c'est quand la nature et l'histoire d'un paradoxe n'ont plus de secret que sa solution n'a plus de secret. Reste que la solution proposée se détourne de l'épistémologie formelle qui a jusque-là entretenu le paradoxe pour entrer dans la métaphysique, et, en abaissant la tension ontique-épistémique, semble ne pas beaucoup renseigner sur ce que nous avons appelé le « Paradoxe suprême ».

Oui, la métaphysique est la branche sur laquelle nous nous accrochons pour sortir de l'enlisement. Remarquons qu'Arnold Zuboff lui-même a tenté, comme il l'écrit, une « application de la probabilité à la

métaphysique »[190] pour résoudre la Belle. Sa méthode est extrême puisqu'il est conduit à penser qu'au-delà de toute fission des individus dans le temps ou dans l'espace, se conserve un même « Je », qu'au-delà de la pluralité et de la variété des consciences qui peuplent l'espèce humaine et éventuellement les espèces extraterrestres, se conserve aussi un même être indivis, sur un plan d'existence « plus vrai » que le plan multiple. Sans vouloir déformer la pensée de ce philosophe, nous pensons qu'elle confine à un mysticisme qui fait de l'unité un bien, de la multiplicité un mal. On ne peut pas dire qu'il soit beaucoup suivi sur cette voie. Nous pensons dépasser son intuition selon laquelle les problèmes d'auto-localisation sont aussi des problèmes métaphysiques, justement parce que nous en faisons *in fine* des problèmes métaphysiques. Par exemple, osciller préalablement entre 1/2 et 1/3, encore pris dans les griffes d'un paradoxe *probabiliste*, c'est éviter plus tard de fuir le plan de la multiplicité (du 1/3) et d'être ébloui par la seule unité pourtant immanente (qui voit 1/2), attitude jugée fort prudente par les meilleurs penseurs de semblables paires d'« opposés absolus ». La phrase précédente déconcerte-t-elle ? Nous convenons bien entendu que ce que nous appelons « résoudre le paradoxe » est inévitablement un engagement vers des problèmes d'un autre genre, nous décevons ainsi ceux qui ne juraient que par l'épistémologie formelle. C'est pourtant notre voie, elle sera poursuivie.

Non, le Paradoxe suprême de la probabilité une et deux n'est pas vaincu, loin de là. Mais nous avons appris beaucoup. Nous avons surtout appris qu'on ne lui échappe pas en tranchant le lien ontique-épistémique. Certes, sur la pente de la radicalisation, un subjectiviste peut être un simple demiste refusant de calibrer un crédit sur la chance. Nous pensons maintenant qu'un tel degré de croyance n'est pas ce qu'il convient d'appeler une probabilité. Nous estimons aussi qu'une fréquence et une

[190] Zuboff (2009), p. iii.

propension, des propriétés très liées en réalité, ne sont rien de pertinent pour la science si elles ne se muent pas en crédit. La probabilité n'est pas accidentellement hybride, elle l'est essentiellement. Il faut encourager les approches allant dans ce sens, le bayésianisme objectiviste par exemple, ou une théorie future encore plus satisfaisante.

# Bibliographie


ARMATTE Michel, 2011, « Les marches de l'aléa », *Prisme*, *21*, Centre Cournot pour la Recherche en Économie.

ARNTZENIUS Frank :

- 2002, « Reflections on Sleeping Beauty », *Analysis*, *62 (1)*, p. 53-62.
- 2003, « Some Problems for Conditionalization and Reflection », *The Journal of Philosophy*, *100 (7)*, p. 356-370.

BARBEROUSSE Anouk, 2000, *La physique face à la probabilité*, coll. « Mathesis », Paris, Librairie philosophique J. Vrin.

BARRAU Aurélien, 2007, *Quelques éléments de physique et de philosophie des multivers*, en ligne sur le site du Laboratoire de physique subatomique et de cosmologie de Grenoble : http://lpsc.in2p3.fr/barrau/aurelien/multivers_lpsc.pdf

BARROW John D., TIPLER Frank J., 1986, *The Anthropic Cosmological Principle*, Oxford University Press.

BARTHA Paul, 2006, « How to Put Self-Locating Information in its Place », PhilSci Archive 2993.

BERTRAND Joseph, 1889, *Calcul des probabilités*, Paris, Gauthier-Villars et fils.

BIENVENU Alexis, 2007, *Un empirisme risqué : la philosophie des probabilités de Hans Reichenbach*, thèse dirigée par Michel Bitbol, université Paris 1 Panthéon-Sorbonne.

BOSTROM Nick :

- 2000, « Observer-relative chances in anthropic reasoning? », *Erkenntnis*, *52 (1)*, p. 93-108.
- 2002, *Anthropic Bias: Observation Selection Effects in Science and Philosophy*, coll. « Studies in Philosophy », New York/Londres, Routledge.
- 2007, « Sleeping Beauty and Self-Location: A Hybrid Model », *Synthese*, *157 (1)*, p. 59-78.

BRADLEY Darren J. :

- 2003, « Sleeping Beauty: A Note on Dorr's Argument for 1/3 », *Analysis*, *63 (3)*, p. 266-268.
- 2007, *Bayesianism and self-locating beliefs* or *Tom Bayes meets John Perry*, dissertation doctorale dirigée par Elliott Sober et Michael Friedman, Stanford University.
- 2011a, « Confirmation in a Branching World: The Everett Interpretation and Sleeping Beauty », *British Journal for the Philosophy of Science*, *62 (2)*, p. 323-342.
- 2011b, « Self-location is No Problem for Conditionalization », *Synthese*, *182 (3)*, p. 393-411.
- 2015, « Everettian Confirmation and Sleeping Beauty: Reply to Wilson », *British Journal for the Philosophy of Science*, *66 (3)*, p. 683-693.

BRADLEY Darren J., LEITGEB Hannes, 2006, « When Betting Odds and Credences Come Apart: More Worries for Dutch Book Arguments », *Analysis*, *66 (2)*, p. 119-127.

BRU Bernard, 1994, « Condorcet, mathématique sociale et vérité », *Mathématiques et sciences humaines*, *128*, p. 5-14.

CARNAP Rudolf, 2015, *Logique inductive et probabilité. 1945-1970*, textes traduits sous la direction de Pierre Wagner, coll. « Mathesis », Paris, Librairie philosophique J. Vrin.

CARTER Brandon :

- 1974, « Large number coincidences and the anthropic principle in cosmology », dans Malcolm S. Longair, *International Astronomical Union Symposium No. 63: Confrontation of Cosmological Theories with Observational Data*, Dordrecht, D. Reidel Publishing Company, p. 291-298.
- 2006, « Anthropic Principle in Cosmology », arXiv:gr-qc/0606117.

COURNOT Antoine Augustin, 1843, *Exposition de la théorie des chances et des probabilités*, Paris, Hachette.

COZIC Mikaël, 2007, « Imaging and Sleeping Beauty: A Case for Double Halfers », dans D. Samet, *Proceedings of TARK 2007 Conference*, Presses Universitaires de Louvain, p. 112-117.

COZIC Mikaël, DROUET Isabelle, 2009, « Interpréter les probabilités », *Pour la science*, *385*, p. 52-58.

COZIC Mikaël, WALLISER Bernard, 2012, « Ce que mesurent les probabilités », *Prisme*, *24*, Centre Cournot pour la Recherche en Économie.

CRÉPEL Pierre, 2009, « Doutes sur les probabilités », *Les Génies de la science*, *39*, p. 62-69.

D'ALEMBERT Jean le Rond :

- 1754, « Croix ou pile », dans Denis Diderot, *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, tome 4, Paris, Briasson, David, Le Breton et Durand, p. 512-513.
- 1761, *Opuscules mathématiques ou Mémoires sur différents sujets de géométrie, de mécanique, d'optique, d'astronomie, etc.*, vol. 2, Paris, David.

DELABRE Laurent, 2015, « Un jeune paradoxe : la Belle au bois dormant », *Implications Philosophiques*, en ligne : http://www.implications-philosophiques.org/implications-epistemologiques/sciences/paradoxes/belleauboisdormant/

DELABRE Laurent, GERVILLE-RÉACHE Léo, 2015, « Insaisissable Belle au bois dormant », *Philosophia Scientiae*, *19 (1)*, p. 251-269.

DELAHAYE Jean-Paul, 2003, « La Belle au bois dormant, la fin du monde et les extraterrestres », *Pour la Science*, *309*, p. 98-103.

DELANNOY Henri Auguste, 1895, « Sur une question de probabilités traitée par d'Alembert », *Bulletin de la Société Mathématique de France*, *23*, p. 262-265.

DIDEROT Denis, 1761, « Sur deux mémoires de d'Alembert, l'un concernant le calcul des probabilités, l'autre l'inoculation », dans *Œuvres complètes*, éd. J. Assézat et M. Tourneux, vol. 9, Paris, Garnier, 1875, p. 192-212.

DORR Cian, 2002, « Sleeping Beauty: in defence of Elga », *Analysis*, *62 (4)*, p. 292-296.

DRAPER Kai, 2017, « Even for objectivists, Sleeping Beauty isn't so simple », *Analysis*, *77 (1)*, p. 29-37.

ELGA Adam, 2000, « Self-Locating Belief and the Sleeping Beauty Problem », *Analysis*, *60 (2)*, p. 143-147.

ENGEL Pascal, 1995, « Le fidéisme de van Fraassen », conférence à l'université du Québec à Montréal, ms. non publié, en ligne sur le site de l'université de Genève : https://www.unige.ch/lettres/philo/enseignants/pe/Engel%201995%20Le%20fideisme%20de%20Van%20Fraassen.pdf

FRANCESCHI Paul :

- 2005, « Sleeping Beauty and the Problem of World Reduction », PhilSci Archive 2175.
- 2010, « A Two-Sided Ontological Solution to the Sleeping Beauty Problem », PhilSci Archive 8357.

FREGE Gottlob, 1971, *Écrits logiques et philosophiques*, trad. Claude Imbert, coll. « Points - Essais », Paris, Seuil.

GÄRDENFORS Peter, 1988, *Knowledge in flux. Modeling the dynamics of epistemic states*, Bradford Books, Cambridge (Mass.), The MIT Press.

GERVILLE-RÉACHE Léo, 2016, « Le non-paradoxe du conducteur distrait », *Implications Philosophiques*, en ligne : http://www.implications-philosophiques.org/implications-epistemologiques/sciences/paradoxes/le-non-paradoxe-du-conducteur-distrait/

GILLIES Donald, 2000, *Philosophical Theories of Probability*, Londres/New York, Routledge.

GROISMAN Berry, 2008, « The End of Sleeping Beauty's Nightmare », *British Journal for the Philosophy of Science*, *59 (3)*, p. 409-416.

HACKING Ian MacDougall :

- 1971, « Equipossibility Theories of Probability », *British Journal for the Philosophy of Science*, *22 (4)*, p. 339-355.
- 1993, *Le plus pur nominalisme. L'énigme de Goodman : « vleu » et usages de « vleu »*, trad. R. Pouivet, Combas, Éditions de l'Éclat.
- 2002, *L'émergence de la probabilité*, trad. M. Dufour, coll. « Liber », Paris, Seuil.

HITCHCOCK Christopher Read, 2004, « Beauty and the Bets », *Synthese*, *139 (3)*, p. 405-420.

HORGAN Terence Edward :

- 2004, « Sleeping Beauty Awakened: New Odds at the Dawn of the New Day », *Analysis*, *64 (1)*, p. 10-21.
- 2007, « Synchronic Bayesian Updating and the Generalized Sleeping Beauty Problem », *Analysis*, *67 (1)*, p. 50-59.
- 2008, « Synchronic Bayesian Updating and the Sleeping Beauty Problem: Reply to Pust », *Synthese*, *160 (2)*, p. 155-159.

HORGAN Terence Edward, MAHTANI Anna, 2013, « Generalized Conditionalization and the Sleeping Beauty Problem », *Erkenntnis*, *78 (2)*, p. 333-351.

JACKSON Frank, PRIEST Graham (dir.), 2004, *Lewisian Themes: The Philosophy of David K. Lewis*, Oxford University Press (Clarendon Press).

KATSUNO Hirofumi, MENDELZON Alberto O., 1992, « On the difference between updating a knowledge base and revising it », dans Peter Gärdenfors, *Belief Revision*, Cambridge University Press, p. 183-203.

LESLIE John Andrew :

- 1987, « Probabilistic phase transitions and the anthropic principle », dans J. Demaret, *Origin and Early History of the Universe: Proceedings of the 26th Liège International Astrophysical Colloquium*, Cointe-Ougrée, Belgique, Université de Liège, Institut d'astrophysique, p. 439-444.
- 1992, « Time and the Anthropic Principle », *Mind*, *101 (403)*, p. 521-540.
- 1996, *The End of the World: The Science and Ethics of Human Extinction*, Londres/New York, Routledge.

LEVI Isaac, 1994, « Changing Probability Judgements », dans Paul Humphreys, *Patrick Suppes: Scientific Philosopher*, *Volume 1: Probability and Probabilistic Causality*, Kluwer Academic Publishers, p. 87-108.

LEWIS David Kellogg :

- 1976, « Probabilities of Conditionals and Conditional Probabilities », *Philosophical Review*, *85 (3)*, p. 297-315.
- 1979, « Attitudes *De Dicto* and *De Se* », *Philosophical Review*, *88 (4)*, p. 513-543.
- 1980, « A Subjectivist Guide to Objective Chance », dans Richard C. Jeffrey, *Studies in Inductive Logic and Probability*, vol. 2, Berkeley, University of California Press, p. 263-293.
- 1994, « Humean Supervenience Debugged », *Mind*, *103 (412)*, p. 473-490.
- 2001, « Sleeping Beauty: Reply to Elga », *Analysis*, *61 (3)*, p. 171-176.
- 2007, *De la pluralité des mondes*, trad. M. Caveribère et J.-P. Cometti, Combas, Éditions de l'Éclat.

LEWIS Peter J. :

- 2007, « Quantum Sleeping Beauty », *Analysis*, *67 (1)*, p. 59-65.
- 2009, « Reply to Papineau and Durà-Vilà », *Analysis*, *69 (1)*, p. 86-89.
- 2010, « Credence and self-location », *Synthese*, *175 (3)*, p. 369-382.

MARTIN Thierry :

- 1994, « La valeur objective du calcul des probabilités selon Cournot », *Mathématiques et sciences humaines*, *127*, p. 5-17.
- 1996, *Probabilités et critique philosophique selon Cournot*, coll. « Mathesis », Paris, Librairie philosophique J. Vrin.

MEACHAM Christopher J. G. :

- 2008, « Sleeping Beauty and the Dynamics of De Se Beliefs », *Philosophical Studies*, *138 (2)*, p. 245-269.
- 2010, « Two Mistakes Regarding the Principal Principle », *British Journal for the Philosophy of Science*, *61 (2)*, p. 407-431.

MELLOR David Hugh, 1971, *The Matter of Chance*, Cambridge University Press.

MILLER David W., 1966, « A Paradox of Information », *British Journal for the Philosophy of Science*, *17 (1)*, p. 59-61.

MONTON Bradley, 2002, « Sleeping Beauty and the Forgetful Bayesian », *Analysis*, *62 (1)*, p. 47-53.

MONTON Bradley, KIERLAND Brian, 2005, « Minimizing Inaccuracy for Self-Locating Beliefs », *Philosophy and Phenomenological Research*, *70 (2)*, p. 384-395.

MOSTERÍN Jesús, 2004, « Anthropic Explanations in Cosmology », dans Petr Hájek, Luis Valdés-Villanueva, Dag Westerståhl, *Proceedings of the 12th International Congress of Logic, Methodology and Philosophy of Science*, Amsterdam, North-Holland Publishing, p. 441-471.

NEAL Radford M., 2006, *Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning*, Technical Report No. 0607, Department of Statistics, University of Toronto, arXiv:math/0608592.

PAPINEAU David, DURÀ-VILÀ Victor :

- 2009a, « A thirder and an Everettian: A reply to Lewis's 'Quantum Sleeping Beauty' », *Analysis*, *69 (1)*, p. 78-86.
- 2009b, « Reply to Lewis: metaphysics versus epistemology », *Analysis*, *69 (1)*, p. 89-91.

PAQUETTE Michel, 2002, « L'explication du choix rationnel chez Carnap », dans F. Lepage, M. Paquette et F. Rivenc, *Carnap aujourd'hui*, coll. « Analytiques », Montréal/Paris, Bellarmin/Vrin, p. 59-86.

PASCAL Blaise, 1963, *Œuvres complètes*, éd. L. Lafuma, Paris, Seuil.

PATY Michel, 1988, « D'Alembert et les probabilités », dans Roshdi Rashed, *Sciences à l'époque de la Révolution française. Recherches historiques*, Paris, Librairie Blanchard, p. 203-265.

PERRY John, 1993, *The Problem of the Essential Indexical and Other Essays*, Oxford University Press.

PETTIGREW Richard, 2012, « Accuracy, Chance, and the Principal Principle », *Philosophical Review*, *121 (2)*, p. 241-275.

PICCIONE Michele, RUBINSTEIN Ariel, 1997, « On the Interpretation of Decision Problems with Imperfect Recall », *Games and Economic Behavior*, *20 (1)*, p. 3-24.

POINCARÉ Henri, 2013, *La science selon Henri Poincaré : La science et l'hypothèse, La valeur et la science, Science et méthode*, textes présentés par Jean-Pierre Bourguignon, coll. « Idem », Paris, Dunod.

POLLOCK John Leslie *et al.* (The OSCAR Seminar), 2008, « An Objectivist Argument for Thirdism », *Analysis*, *68 (2)*, p. 149-155.

PUST Joel :

- 2008, « Horgan on Sleeping Beauty », *Synthese*, *160 (1)*, p. 97-101.
- 2011, « Sleeping Beauty and direct inference », *Analysis*, *71 (2)*, p. 290-293.
- 2012, « Conditionalization and Essentially Indexical Credence », *The Journal of Philosophy*, *109 (4)*, p. 295-315.
- 2013, « Sleeping Beauty, evidential support and indexical knowledge: reply to Horgan », *Synthese*, *190 (9)*, p. 1489-1501.

QUINE Willard Van Orman, 1969, « Propositional Objects », dans *Ontological Relativity and Other Essays*, Columbia University Press, p. 139-160.

SCHMITT Yann (dir.), 2012, *La philosophie de David Lewis*, *Klésis Revue philosophique*, *24*.

SCHWARZ Wolfgang, 2015, « Lost memories and useless coins: revisiting the absentminded driver », *Synthese*, *192 (9)*, p. 3011-3036.

STALNAKER Robert C., 2008, *Our Knowledge of the Internal World*, coll. « Lines of Thought », Oxford University Press (Clarendon Press).

STREVENS Michael, 1995, « A Closer Look at the 'New' Principle », *British Journal for the Philosophy of Science*, *46 (4)*, p. 545-561.

TATIN-GOURIER Jean-Jacques (éd.), 1994, *Cagliostro et l'affaire du Collier. Pamphlets et polémiques*, Publications de l'Université de Saint-Étienne.

THORN Paul D. :

- 2011, « Undercutting defeat via reference properties of differing arity: a reply to Pust », *Analysis*, *71 (4)*, p. 662-667.
- 2012, « Two Problems of Direct Inference », *Erkenntnis*, *76 (3)*, p. 299-318.

TITELBAUM Michael G. :

- 2008, « The Relevance of Self-Locating Beliefs », *Philosophical Review*, *117 (4)*, p. 555-605.
- 2012, « An Embarrassment for Double-Halfers », *Thought: A Journal of Philosophy*, *1 (2)*, p. 146-151.
- 2013, *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*, Oxford University Press.

VAIDMAN Lev, SAUNDERS Simon, 2001, « On Sleeping Beauty Controversy », PhilSci Archive 324.

VAN FRAASSEN Bas C. :

- 1984, « Belief and the Will », *The Journal of Philosophy*, *81 (5)*, p. 235-256.
- 1995, « Belief and the Problem of Ulysses and the Sirens », *Philosophical Studies*, *77 (1)*, p. 7-37.
- 1999, « Conditionalization: a new argument for », *Topoi*, *18 (2)*, p. 93-96.

WALLISER Bernard, BARATGIN Jean, 2010, « Sleeping Beauty and the absent-minded driver », *Theory and Decision*, *69 (3)*, p. 489-496.

WALLISER Bernard, ZWIRN Denis, 2002, « Can Bayes' rule be justified by cognitive rationality principles? », *Theory and Decision*, *53 (2)*, p. 95-135.

WEATHERSON Brian, 2011, « Stalnaker on Sleeping Beauty », *Philosophical Studies*, *155 (3)*, p. 445-456.

WEINTRAUB Ruth, 2004, « Sleeping Beauty: a simple solution », *Analysis*, *64 (1)*, p. 8-10.

WEISBERG Jonathan, 2007, « Conditionalization, Reflection, and Self-Knowledge », *Philosophical Studies*, *135 (2)*, p. 179-197.

WHITE Roger, 2006, « The Generalized Sleeping Beauty Problem: A Challenge for Thirders », *Analysis*, *66 (2)*, p. 114-119.

WILSON Alastair, 2014, « Everettian Confirmation and Sleeping Beauty », *British Journal for the Philosophy of Science*, *65 (3)*, p. 573-598.

ZUBOFF Arnold :

- 1990, « One Self: The Logic of Experience », *Inquiry*, *33 (1)*, p. 39-68.
- 2009, *Time, Self and Sleeping Beauty*, dissertation doctorale dirigée par Thomas Nagel, Princeton University.

# Table des matières

# L'interprétation double de la probabilité et les problèmes d'auto-localisation

( *The dual interpretation of probability and the problems of self-location* )

L'histoire de la théorie de la probabilité peut être perçue comme l'évolution des relations entre deux interprétations, épistémique et ontique. Certains problèmes récents décrivent la situation d'un agent rationnel aux croyances auto-localisantes : il estime la probabilité de ses possibles positions dans le monde (temps, espace, identité) et plus seulement dans l'espace logique. Il arrive qu'un fréquentiste ou un adepte d'une lecture ontologique de la probabilité, en essayant de se localiser, conclut des résultats si différents de ceux du bayésien plus subjectif, que se crée un inédit et inquiétant écart entre leurs probabilités. Démêler cette situation paradoxale passe par une enquête sur les théories philosophiques et scientifiques qui l'ont engendrée, dont la réunion chez d'importants penseurs contemporains comme David Lewis est peut-être l'occasion d'un retour, sous une forme neuve plus acceptable, des intuitions incomprises de Jean le Rond D'Alembert. Cette recherche de liens historiques entre interprétation double et auto-localisation mène à l'analyse de plusieurs problèmes et surtout à l'étude du récent mais déjà célèbre paradoxe de la Belle au bois dormant : les principales tentatives de résolution trouvées dans la littérature sont longuement expliquées et critiquées, puis est esquissée une solution originale qui prétend réparer la scission ontique-épistémique due à la perte d'un repère temporel de l'agent rationnel.

***Mots-clés*** : auto-localisation, bayésianisme, Belle au bois dormant, conditionalisation, D'Alembert, décision, dynamique des croyances, fréquentisme, identité personnelle, indexicalité, Lewis David, mémoire, mondes possibles, paradoxe, principes d'alignement, probabilité épistémique, probabilité ontique, raisonnement anthropique, temps

Université Paris 1 Panthéon-Sorbonne

UFR de Philosophie
17, rue de la Sorbonne
75005 Paris

É. D. de Philosophie
1, rue d'Ulm
75005 Paris

IHPST
13, rue du Four
75006 Paris