La théorie argumentative du raisonnement

Hugo Mercier

Hugo MERCIER

La Théorie argumentative du raisonnement

ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES

Thèse de doctorat en sciences sociales, mention sciences cognitives

Dirigée par Dan SPERBER

Date de soutenance : 5 janvier 2009

Membres du Jury

Daniel Andler (Université Paris-Sorbonne)
Didier Bazalgette (Délégation Générale pour l'Armement)
Jérôme Dokic (Ecole des Hautes Etudes en Sciences Sociales)
Ira Noveck (Centre National de la Recherche Scientifique, rapporteur)
Guy Politzer (Centre National de la Recherche Scientifique, rapporteur)
Dan Sperber (Centre National de la Recherche Scientifique, directeur de thèse)
Jean-Baptiste Van der Henst (Centre National de la Recherche Scientifique)

Table des matières

Table des matières détaillée

Remerciements

Mes remerciements s'adressent tout d'abord à Dan Sperber pour avoir inspiré puis supervisé cette thèse. Des attentes si hautes que les miennes avant ce travail sont généralement déçues. Elles se sont avérées trop basses. Je souhaite de tout cœur que nous pourrons poursuivre une collaboration qui a été pour moi aussi enrichissante intellectuellement que personnellement.

Jean-Baptiste Van der Henst et Guy Politzer m'ont fournit de précieux conseils tout au long de cette thèse, ce pour quoi je leur suis très reconnaissant. Je les remercie également d'avoir accepté de faire partie de mon jury de thèse. Jean-Baptiste et moi avons été de nombreuses expéditions ensemble – de Stresa à Kobe – et j'espère que nos pourrons continuer nos explorations interculturelles.

C'est grâce à un cours d'Ira Noveck que je suis ici maintenant, grâce à ce cours stimulant et à l'incroyable accueil qu'il avait offert dans son équipe à un tout jeune étudiant. Il sera là aussi à la fin de ma vie d'étudiant en ayant accepté de faire partie du jury.

Je remercie également Jérôme Dokic et Daniel Andler d'avoir accepté de faire partie du jury, j'attends leurs commentaires avec impatience.

La DGA m'a financé durant cette thèse, et je tiens en particulier à remercier Didier Bazalgette, qui a également accepté de faire partie du jury.

J'ai joui durant cette thèse de l'entourage des membres du NaSH (élargi), Jean-Baptiste André, Nicolas Baumard, Coralie Chevallier, Nicolas Claidière, Christophe Heintz, Yasmina Jraissati, Olivier Mascaro, Olivier Morin, Aniko Sebesteny et Hugo Viciana. Tout ceci aurait été assez ennuyeux sans eux. J'espère que nous resterons collègues, et surtout amis, pendant longtemps ! Avec tous les amis qui ont contribué à faire de ces années à Paris des moments si riches et plaisants, vous allez me manquer.

Finalement, et surtout, je remercie ma famille pour son soutien indéfectible durant mes errements d'étudiant.

Just consider how terrible the day of your death will be.
Others will go on speaking, and you will not be able to argue back.
Ram Mohun Roy

Introduction

> La démocratie et l'aristocratie ne sont point des Etats libres par leur nature. La liberté politique ne se trouve que dans les gouvernements modérés. Mais elle n'est pas toujours dans les États modérés. Elle n'y est que lorsqu'on n'abuse pas du pouvoir : mais c'est une expérience éternelle que tout homme qui a du pouvoir est porté à en abuser ; il va jusqu'à ce qu'il trouve des limites. Qui le dirait ! la vertu même a besoin de limites.
>
> Pour qu'on ne puisse abuser du pouvoir, il faut que, par la disposition des choses, le pouvoir arrête le pouvoir.
>
> Montesquieu, De l'esprit des lois, Livre XI, Chapitre IV

C'est ainsi que Montesquieu cherche à convaincre son lecteur de la nécessité de la division des pouvoirs. Si une personne est dépositaire du pouvoir elle ne pourra qu'en abuser, il est donc nécessaire que d'autres pouvoirs servent de contrepoint. C'est notre faculté de raison qui nous permet de saisir cet argument ô combien important. Mais on aurait tout aussi bien pu choisir un exemple plus prosaïque. « C'est moi qui ai fait la vaisselle hier, c'est ton tour aujourd'hui. » De même qu'un roi, contraint par sa raison, doit céder aux arguments de Montesquieu, nous devons nous soumettre à la logique imparable de notre partenaire. Bien que nous utilisions tous, quotidiennement, nos capacités de raisonnement pour argumenter, on peut être tenté de penser qu'il ne s'agit là, pour ce sommet de notre intellect, que d'une tâche presque subalterne : le raisonnement ne nous permet-il pas aussi, et d'abord, de comprendre le monde, d'accéder à des vérités supérieures, d'appréhender la logique ou les mathématiques, et de résoudre des problèmes aussi divers que la preuve du théorème de Fermat et le mystère du meurtre de l'Orient Express ? Que le raisonnement nous permette de faire toutes ces choses est indéniable. Que ce soit là sa *fonction* est une autre question. Il ne faut pas se laisser abuser par l'importance ou la beauté des accomplissements rendus possibles par le raisonnement. Ce n'est pas

parce que les mains ont permis la création de La Vénus de Milo, de Notre Dame de Paris ou du boulier que leur fonction est de réaliser de merveilleux artefacts. L'objectif de cette thèse sera de montrer que le raisonnement n'est *pas fait* pour résoudre des problèmes ou prendre de meilleures décisions, mais au contraire pour *argumenter*, pour trouver des arguments convaincants et les évaluer ; je défendrai donc une *théorie argumentative du raisonnement* (Mercier & Sperber, in press; Sperber, 2000, 2001; Sperber & Mercier, in prep, voir aussi Dessalles, 2000).

De façon bien compréhensible, les psychologues du raisonnement voient son œuvre partout. Ainsi, pour Johnson-Laird et Byrne, la déduction sert

> to formulate plans and to evaluate actions; to determine the consequences of assumptions and hypotheses; to interpret and to formulate instructions, rules, and general principles; to pursue arguments and negotiations; to weigh evidence and to assess data; to decide between competing theories; and to solve problems (Johnson-Laird & Byrne, 1991, p.3)

De même, pour Rips, « deductive reasoning […] is cognitively central » (Rips, 1994, p.11), et il illustre le rôle central de la déduction par des exemples montrant son importance pour la compréhension et la planification. Bien qu'ils démontrent en effet l'utilité que peut avoir la déduction dans ces processus, ces exemples peuvent également être interprétés comme en montrant les limites. Rips reprend ainsi l'échange suivant (Ibid., p.13) :

> A: Will Burke win?
> B: He's the machine candidate, and the machine candidate always wins

Pour comprendre que B répond bien à la question de A, il faut utiliser le raisonnement – déductif dans ce cas – pour conclure que Burke va en effet gagner. Mais de nombreux autres processus sont à l'œuvre qui nous permettent la compréhension de cette phrase, processus phonétiques, phonologiques, syntaxiques, sémantiques et pragmatiques. Si le raisonnement est bien essentiel pour comprendre *cette* phrase, il ne le serait sans doute pas pour comprendre celle-ci :

B: Yes, he's going to win

Bien que la compréhension de cet énoncé paraisse triviale, elle est en fait très complexe : il faut inférer que le 'Yes' réfère à l'énoncé précédent, que 'he' est Burke, et que ce qu'il va gagner est la chose mentionnée dans la question (sans mentionner les autres étapes de traitement de la phonétique à la sémantique). Que nous ne nous rendions même pas compte de leur présence ne fait que révéler leur extraordinaire efficacité, qui est elle-même un signe de leur importance.

Il est possible de faire une analyse similaire de l'exemple que Rips donne de planification. Gary veut voir un film et hésite entre deux choix :

> Either I will go to *Hiroshima, mon amour*, or I will go to *Duck Soup*.
> I won't go to *Hiroshima, mon amour*.
> (therefore) I will go to *Duck Soup*. (Ibid., p.13)

Dans ce cas, Gary utilise un syllogisme disjonctif pour faire son choix. Là encore, on peut considérer qu'une forme de déduction a eu lieu. Mais quelle part joue-t-elle vraiment dans le processus de planification, ou de prise de décision ? Il est clair que de nombreux autres mécanismes, exécutifs, moteurs, perceptifs, mémoriels, dont nous sommes à peine – ou pas du tout – conscients vont jouer un rôle encore plus important.

Ce que ces deux exemples démontrent c'est que le raisonnement, s'il joue un rôle dans ces divers processus de prise de décision ou de compréhension, est loin d'être le seul mécanisme à l'œuvre. Une première tâche, avant même d'essayer de comprendre comment fonctionne le raisonnement, est d'essayer de circonscrire son activité. Par exemple, Johnson-Laird et Byrne divisent les processus de pensée en « association, création, induction, déduction et calcul » (1991, p.2), restreignant le terme de raisonnement aux trois derniers. Plus récemment, Johnson-Laird a proposé de diviser les processus mentaux en catégories basées sur l'accès conscient à leurs prémisses et à leurs conclusions (Johnson-Laird, 2006, p.60ff), mais les limites du terme 'raisonnement' ne sont pas claires. De même, Rips précise bien que :

> not all mental transformations are deductive ones, and we need to leave room for obviously nondeductive transformations that can occur in image

manipulation, in forgetting, and in various sorts of inductive and analogical inference. (Rips, 1994, pp.10-11).

On voit donc que les psychologues du raisonnement sont bien conscients des limites de leur objet d'étude, et du besoin de postuler l'existence d'autres types de mécanismes psychologiques. Ce sont ces observations qui seront à la base de la création des *théories à processus duel*.

Comme nous le verrons dans le chapitre qui leur est consacré, ces théories à processus duel restent généralement vagues dans leur caractérisation des mécanismes de raisonnement par opposition aux autres mécanismes psychologiques. Dans le cadre de la théorie argumentative, cette division devient une différence entre processus d'inférences *intuitifs* et *réflexifs* (Mercier & Sperber, in press; Sperber & Mercier, in prep). Inférence est ici entendue en son sens le plus général de processus admettant un input, le traitant, et fournissant un output informationnellement enrichi. Il peut donc s'agir tout aussi bien de traitements dans les rouages du système visuel (Kersten, Mamassian, & Yuille, 2004) que des formes les plus conscientes, explicites de raisonnement. Ce qui distingue les inférences réflexives des inférences intuitives est le fait qu'elles traitent de *raisons* : elles examinent des représentations pour déterminer s'il s'agit de bonnes raisons pour faire une inférence donnée. Il s'agit d'un système métareprésentationnel qu'on peut, car il traite justement de raisons, qualifier de *raisonnement*. Par opposition, on peut simplement qualifier les inférences intuitives d'*intuitions*. C'est le raisonnement qui nous a servi lorsque nous avons examiné le lien entre la prémisse « Si une personne est dépositaire du pouvoir elle ne pourra qu'en abuser » et la conclusion « il est donc nécessaire que d'autres pouvoirs servent de contrepoint » : dans ce cas, la prémisse est examinée en tant que représentation et, plus spécifiquement, ce sont ses propriétés en tant que raison soutenant la conclusion qui sont traitées. Il est clair qu'il ne peut s'agir là que d'un système métareprésentationnel : les objets du monde (autres que des représentations) ne peuvent pas être de bonnes ou de mauvaises raisons, ne peuvent pas être des raisons tout court.

Il s'agit là d'une esquisse (une version plus détaillée sera présentée dans les chapitres 2 et 3) de la solution que la théorie argumentative propose pour circonscrire le raisonnement. Mais borner les mécanismes de raisonnement n'est pas le seul problème auquel sont confrontés les psychologues traitant du sujet. Depuis le

renouveau connu par l'étude empirique du raisonnement dans les années 60, l'image qui se dessine de nos capacités de raisonnement ne prête guère à l'optimisme. De même que les chercheurs étudiant la prise de décision, les psychologues du raisonnement ont catalogué erreurs, biais et approximations. Face à ce type de résultats, il y a deux solutions : remettre en cause la fonction attribuée à la capacité testée, ce qu'on pense qu'elle est censée faire, ou rendre compte des résultats par des contraintes sur ce système et l'intervention malvenue d'autres systèmes. C'est uniformément la seconde solution qu'ont choisie les psychologues du raisonnement, blâmant les limites de la mémoire de travail et les mécanismes intuitifs pour les mauvaises performances du raisonnement. Dans cette thèse je vais explorer la solution alternative : remettre en cause la fonction du raisonnement.

Les tâches utilisées par les psychologues du raisonnement sont très artificielles. Les participants peuvent par exemple être confrontés aux prémisses suivantes :

Aucun boulanger n'est un pianiste
Certains pianistes sont des footballeurs

Et ils doivent déterminer s'il est possible de dériver une conclusion logiquement valide de ces prémisses et, si oui, laquelle[1]. Dans ce cas, on peut déduire des prémisses que certains footballeurs ne sont pas des boulangers. A peine plus d'un quart des participants est généralement capable de déterminer que cette conclusion est la bonne (voir Geurts, 2003). Les mauvaises performances dans ce type de tâche sont parfois expliquées par la mémoire de travail, dont la faible capacité empêcheraient les participants d'y maintenir un modèle suffisamment complexe des prémisses (Johnson-Laird & Byrne, 1991). Une autre source de confusion possible, pour les participants, serait introduite par des mécanismes intuitifs de pragmatique qui les entraîneraient à conclure, de 'certains pianistes sont des footballeurs', que 'tous les pianistes ne sont pas des footballeurs'. Or cette inférence ne respecte pas les règles de la logique et pourrait les mettre sur une mauvaise piste (Newstead, 1989, 1995).

[1] Conclusion concernant les footballeurs et les boulangers.

On peut suggérer une analogie pour illustrer les limites de ce type d'explication. Prenez des participants et demandez-leur de marcher sur les mains. A moins que vous ne preniez des experts, leurs performances ne seront sûrement pas très bonnes. Vous pourrez alors expliquer ces piètres résultats par des contraintes (les muscles des bras sont vraiment trop faibles) ou par les effets néfastes d'autres structures (c'est à cause des jambes si les gens n'arrivent pas à marcher sur les mains, elles ne cessent de les déséquilibrer). Une autre explication est que les mains ne sont pas faites pour marcher… Il s'agit dans le premier cas d'une explication proximale, en termes de mécanismes, et dans le second d'une explication ultime, en termes de fonctions (Tinbergen, 1963). Ces deux explications ne sont pas incompatibles : pour qui veut comprendre comment on marche sur les mains, et pourquoi on y arrive mal, elles sont toutes deux importantes – les explications proximales pouvant même être plus importantes. Mais si on cherche à comprendre comment fonctionnent les mains en général, il est bon de s'intéresser d'abord au niveau ultime. De même pour le raisonnement. Plutôt que de continuer à rendre compte des échecs du raisonnement dans les tâches classiques par des contraintes ou l'action de mécanismes intuitifs, une nouvelle hypothèse concernant la fonction du raisonnement va être proposée.

Considérations théoriques

Pour qui veut s'inscrire dans un cadre évolutionniste, la première tâche est de spéculer sur la fonction des mécanismes étudiés. Le premier chapitre sera donc consacré à l'élaboration d'un scénario plausible qui aurait pu mener à l'apparition de nos capacités de raisonnement, scénario qui donnera des éléments du contexte et des pressions de sélection ayant promu ces capacités. Il s'agira là d'une spéculation dont le rôle n'est pas de décrire un passé dont nous savons trop peu, mais est d'ouvrir une perspective nouvelle sur la psychologie du raisonnement en montrant qu'elle n'est pas sans plausibilité. Une fois que la plausibilité évolutionniste de la théorie argumentative aura été établie, la seconde étape consistera à mettre en évidence les prédictions sur le fonctionnement du raisonnement qui découleraient d'une telle théorie. Enfin, dans un troisième chapitre la théorie défendue ici sera comparée aux autres théories à processus duel existant dans le domaine du raisonnement. Cette première partie sera donc surtout constituée de considérations théoriques, considérations qui seront amplement corroborées dans la seconde partie. Cette division se justifie par le caractère partiellement post-hoc de la théorie présentée ici (ce qui s'applique également à la plupart des autres théories à processus duel). Les expériences qui seront citées en sa faveur n'ont pas été conduites afin de prouver tel ou tel aspect de la théorie et, dans de nombreux cas, ces expériences seront utilisées en soutien de plusieurs arguments. Il était dans ces conditions plus simple de rassembler la plupart des considérations théoriques dans une seule partie, avant de passer en revue les nombreux travaux fournissant un soutien à la théorie argumentative et à ses prédictions sur le fonctionnement du raisonnement.

1 Evolution du raisonnement

A quoi sert le raisonnement ? Comme son nom l'indique, la théorie argumentative défend l'idée que sa fonction principale est de produire et d'évaluer des arguments. Cependant, avant de pouvoir examiner les pressions de sélection ayant pu être conduire à l'évolution d'une telle capacité, par nature communicative, il est nécessaire de tracer, dans les grandes lignes, le cadre plus général de l'évolution de la communication.

1.1 L'évolution de la communication

L'évolution de la communication a commencé de poser problème dans les années 70, lorsque les biologistes de l'évolution (en particulier, Hamilton, 1964a, 1964b; Trivers, 1971; Williams, 1966) ont examiné les sources du mécanisme de sélection naturelle et ont rappelé qu'il agissait principalement au niveau des individus et non au niveau des groupes. Jusqu'alors les éthologues pouvaient prétendre que les systèmes de communication permettaient aux animaux de diffuser des informations et d'apporter des bénéfices au groupe ou à l'espèce. Ainsi, un écureuil poussant un cri d'alarme – au risque de sa vie – permettait aux autres membres du groupe de s'échapper. Le problème posé par cette interprétation fut tout d'abord soulevé par Richard Dawkins et John Krebs dans un chapitre de la première édition de Behavioural Ecology: An Evolutionary Approach (1978). Si on se place au niveau des individus, un signal tel qu'un cri d'alarme – qui met en danger son émetteur au profit du récepteur – devient une aberration : les individus qui les produisent devraient avoir moins de chance de se reproduire, et le trait devrait donc s'éteindre.

Il s'agit donc d'expliquer l'existence même des systèmes de communication observés dans la nature. Dans le cas des cris d'alarme, une piste souvent explorée est celle de la sélection de parentèle : si le cri permet de sauver plusieurs enfants, il devient intéressant pour un parent de le pousser. Dans certains groupes d'écureuils par exemple, les femelles sont des membres permanents alors que les mâles ne sont que de passage. Les premières ont donc plus intérêt à pousser des cris d'alarme car il

y a plus de chance que leurs apparentés soient menacés, et c'est en effet ce qui a été observé (Sherman, 1977), confirmant ainsi une prédiction de la sélection de parentèle. Cette explication ne peut cependant pas rendre compte des systèmes de communication que l'on observe entre individus non apparentés. Dans ces cas, pour Dawkins et Krebs l'intérêt de la communication résiderait plutôt dans la possibilité qu'elle offre pour l'émetteur de manipuler le récepteur : l'émetteur pourrait influencer le récepteur – de par les informations qu'il lui transmet – à son propre avantage. C'est par exemple ce que font les papillons vice-rois. Leurs ailes portent des motifs similaires à ceux des papillons monarques, qui sont des papillons toxiques que les oiseaux évitent. Ces motifs permettent aux vice-rois d'être eux aussi évités par les oiseaux qui ont des difficultés à les différencier des monarques.

Il y a cependant une faille dans cet argument, faille que Krebs et Dawkins identifieront eux-mêmes dans la seconde édition de Behavioural Ecology (1984). Si le fait de recevoir des informations d'un autre individu est réellement dommageable aux membres d'une espèce, alors l'évolution les rendra rapidement 'sourds' à ces informations. Ils arrêteront de les percevoir ou d'y prêter attention. Etant donné que les individus qui envoient des signaux doivent eux aussi y trouver leur compte, il en résulte que la communication, pour être évolutionnairement stable, doit être majoritairement 'honnête'. Cela ne signifie pas que tous les signaux doivent refléter parfaitement la réalité, mais simplement qu'en moyenne les bénéfices qu'ils apportent à la fois à l'émetteur et au récepteur doivent être plus importants que les risques et coûts encourus. Dans le cas des monarques et des vice-rois, il est préférable pour les oiseaux de s'abstenir de manger des vice-rois (qui pourtant ne sont pas toxiques) pour éviter d'ingurgiter un monarque. Si le nombre de vice-rois par rapport aux monarques augmentait trop, le système de communication s'écroulerait, les oiseaux ne prêteraient plus attention aux motifs sur les ailes des monarques ou des vice-rois car les vice-rois attrapés compenseraient le risque d'indigestion dû aux quelques monarques restant.

De nombreuses solutions ont depuis été proposées au problème du maintien de l'honnêteté de la communication. Parmi les premières figure la théorie du signal coûteux d'Amotz Zahavi (voir Zahavi & Zahavi, 1997, pour une revue). Cette théorie propose que certains signaux restent honnêtes car leur émission entraîne un coût, coût que ne sont en mesure de payer que ceux qui émettent le signal de façon honnête. Dans ce cas, les récepteurs peuvent se fier au signal pour indiquer que les

émetteurs sont en mesure de payer le coût associé. Par exemple, les mâles d'une certaine espèce de mouche (C*yrtodiopsis dalmanni*) ont les yeux très écartés, ce qui représente un coût certain (ne serait-ce que par l'énergie nécessaire au développement de ces appendices) et les femelles préfèrent les mâles ayant les yeux les plus écartés. Or il a été montré que seuls des mâles bien dotés génétiquement pouvaient maintenir un large écart entre leurs yeux, et ce même dans des conditions environnementales appauvries, et que cette qualité était héréditaire. Les yeux écartés représentent donc un signal honnête parce que coûteux : étant donné que seuls certains mâles peuvent en endurer les coûts, les femelles peuvent se fier à un large écart entre les yeux pour repérer les mâles ayant une bonne qualité génétique (voir Maynard Smith er Harper, 2003, pp.33ff.).

Cependant la communication honnête peut également évoluer en étant à bas coût ('cheap talk'). En s'en tenant aux primates (qui seront pour nous les points de comparaison les plus pertinents), Gouzoules et Gouzoules notent plusieurs autres solutions qui ont été empiriquement testées (Gouzoules & Gouzoules, 2002). Nous avons déjà évoqué une d'entre elle sous la forme de la sélection de parentèle : si émetteur et récepteur partagent suffisamment de gènes, il n'y a pas nécessairement de conflit d'intérêt. Il est également possible que la situation écologique demande une certaine coordination : les groupes ont par exemple intérêt à rester soudés lors des déplacements. Dans la mesure où les individus ont des intérêts en commun, la communication peut faciliter la coordination. C'est ce qui semble être le cas des groupes de tamarins-lion dorés : les cris poussés par leurs membres facilitent leur cohésion (Boinski, Moraes, Kleiman, Dietz, & Baker, 1994). Enfin, certains primates utilisent des stratégies plus sophistiquées pour ajuster leur confiance en différents émetteurs. Les singes vervets sont capables d'émettre plusieurs cris d'alarme en fonction des prédateurs rencontrés. Lorsqu'un de ces cris, enregistré, est diffusé à d'autres vervets dans des circonstances qui ne le justifient pas (en l'absence de prédateur), ces derniers apprennent qu'il ne faut pas s'y fier et arrêtent d'y prêter attention. Cet apprentissage est hautement spécifique : il ne concerne qu'un type de cri chez un individu (Cheney & Seyfarth, 1990; Gouzoules, Gouzoules, & Miller, 1996). Un individu qui voudrait utiliser ces cris à mauvais escient ne pourrait donc le faire qu'un nombre limité de fois, et en payant un prix : après cela, les autres ne se fient plus à lui, même dans des circonstances qui justifierait qu'ils le fassent. Les macaques rhésus, quant à eux, sont capables d'utiliser de manière appropriée le statut

des autres individus et l'histoire des relations qu'ils ont eue avec eux afin d'ajuster leur communication de manière appropriée – à la fois en tant qu'émetteur et que récepteur (Silk, Kaldor, & Boyd, 2000).

Plus généralement, on peut regrouper certaines des techniques utilisées par les récepteurs pour s'assurer de la véracité des signaux envoyés sous le nom de *vigilance épistémique* (Sperber et al., in prep). La communication humaine pose cependant des problèmes particuliers. A la fois l'importance que la communication joue dans notre espèce et son fonctionnement spécifique font qu'elle occupe une place à part parmi les modes de communications animaux. Pour ces raisons, les ancêtres des humains ont dû renforcer des mécanismes de vigilance épistémique préexistants et en développer de nouveaux.

1.2 Mécanismes de vigilance épistémique

Vérification de cohérence

Le mécanisme vérification de cohérence trouve ses sources dans des mécanismes assez anciens, précédant même le besoin de vigilance épistémique. Il fonctionne suivant le principe général suivant : utiliser la cohérence entre une information communiquée et nos propres états mentaux pour savoir s'il convient de l'accepter ou de la refuser. On peut envisager deux formes différentes de ce mécanisme. Une, beaucoup plus ancienne, concerne les intentions et une autre, plus récente, concerne les croyances. La première utilise le fait que des intentions sont en compétition permanente pour la prise de contrôle du système moteur (voir par exemple Redgrave, Prescott & Gurney, 1999; Wolpert & Kawato, 1998). Ce mécanisme peut également servir de filtre au niveau de la communication. Pour cela, il faut que la communication prenne la forme de 'suggestions', similaires à celles que font les chimpanzés par exemple. Lorsqu'un jeune attire l'attention d'un adulte sur lui, puis essaie de lui faire faire ce qu'il veut, on peut dire qu'il 'suggère' une action à l'adulte (Tomasello, Call, Nagell, Olguin, & Carpenter, 1994; Tomasello, Gust, & Frost, 1989). Si l'adulte exécutait automatiquement cette action, il y aurait un danger de manipulation : les jeunes ne manqueraient pas d'en profiter pour leur faire faire

leurs quatre volontés. Rapidement le système s'écroulerait. Mais si nous acceptons le fait que les adultes ne font quasiment rien automatiquement, alors le problème ne se pose pas[2]. Même si l'enfant parvient à activer chez l'adulte l'intention souhaitée (de le prendre sur son dos, de le chatouiller, etc.), cette intention entre en compétition avec les autres intentions de l'adulte. Ce processus de compétition peut permettre au système cognitif de trouver (approximativement) la meilleure chose à faire étant donné les circonstances. Donc si l'adulte a très faim, l'intention de donner de la nourriture à l'enfant sera une des perdantes de la compétition, et il ne se sera pas fait manipuler.[3] Il ne s'agit en fait pas là d'un réel mécanisme de vigilance épistémique, étant donné qu'aucune modification des mécanismes préalables – et dont la fonction est autrement plus générale que celle de la vigilance épistémique – n'est nécessaire, mais d'un précurseur essentiel.

Au fur et à mesure que la communication se développe, elle porte sur des états mentaux qui se trouvent de plus en plus en amont des intentions. Pour les croyances en particulier, ce système ne peut suffire car elles ne se trouvent pas dans une telle situation de compétition directe. En effet, il n'y a pas pour les croyances de goulot d'étranglement tel que celui créé par l'accès au système moteur pour les intentions. Cependant, dans le cas des croyances d'autres systèmes cognitifs peuvent facilement être recrutés pour jouer un rôle similaire à celui des mécanismes de compétition : les mécanismes de mise à jour des croyances. Lorsque des systèmes perceptifs nous informent que le monde a changé, nous devons mettre à jour nos croyances. Ce système permet de trouver en mémoire les informations pertinentes et de les modifier (ou les effacer les cas échéant). Il est possible de réutiliser ce système pour mettre à jour nos croyances lorsqu'on nous communique quelque chose. Nous pouvons ainsi trouver en mémoire les informations pertinentes. Il doit cependant y avoir une différence essentielle. Dans le cas de la perception, la mise à jour est simple : les nouvelles données issues de la perception prennent le pas sur celles présentes en mémoire. Dans le cas du témoignage, la prudence veut que l'inverse se

[2] Nous verrons plus bas (section 3.3) que les mécanismes purement automatiques, ou 'réflexes' sont en fait très rares.

[3] On peut noter qu'on retrouve quelque chose de similaire dans un cas qui peut sembler surprenant, celui des abeilles : elles accordent elles aussi beaucoup plus de poids aux informations venant de leur propre expérience qu'aux informations communiquées par les fameuses danses – alors même qu'il s'agit de sœurs agissant pour un intérêt en grande partie commun (Gruter, Balbuena, & Farina, 2008).

produise : si une information qui m'est communiquée entre en conflit avec ce que je pensais au préalable, il est souvent plus prudent de me fier à ma mémoire. Si ce n'était pas le cas, nous ne serions que trop facilement manipulables. Il y a donc une différence importante entre le cas des intentions et celui des croyances : dans le premier, aucun mécanisme supplémentaire n'est nécessaire alors que dans le second, un nouveau mécanisme qui inverse l'ordre de préférence entre informations nouvelles et informations déjà en mémoire est requis.

Ces deux mécanismes peuvent donc être regroupés sous l'étiquette de vérification de cohérence. La vérification de cohérence nous assure une certaine tranquillité vis-à-vis de l'influence de la communication. Son problème principal vient de ce qu'elle rejette trop d'information. Pour ce qui est des intentions, elle rejette tout ce qui n'a pas un intérêt immédiatement visible, tout ce qui ne passe pas le filtre de la compétition entre les intentions. Par exemple, si un individu me demande à manger, mais que j'ai un peu faim, je vais refuser de lui donner. Mais peut-être que ce don aurait été suivi d'une récompense encore plus généreuse. Le mécanisme de vérification de cohérence ne peut pas prendre en compte ce type de donnée. Pour ce qui est des croyances, le problème se pose lorsque le monde a changé à mon insu. Si j'ai vu Georges près du feu ce matin, et que dans l'après-midi, alors que je suis dans la forêt, on me dit qu'il est parti pêcher, je vais rejeter l'information. Dans ce cas également il eut été préférable que je l'accepte.

Plus généralement, dans le cas des croyances, il pourrait être utile de dépasser le mécanisme de vérification de cohérence lorsque le monde a changé depuis la formation de nos dernières croyances. Ou encore en cas de différences d'expertise, ou de point de vue. Sur la base des mêmes informations, une autre personne a pu tirer des conclusions différentes (soit parce qu'elle avait au préalable des informations différentes, soit parce qu'elle a des mécanismes de traitement de l'information différents, plus ou moins entraînés que les nôtres dans un domaine particulier).

On retrouve les mêmes situations dans le cas des intentions, avec un ajout intéressant : le cas de la coopération. A partir du moment où les membres d'un groupe commencent de coopérer, de très nombreuses possibilités de jeux à somme non nulle (dans lesquels les différents partenaires profitent tous de la participation à une certaine action) s'offrent à eux. Parmi ces jeux, l'immense majorité bénéficiera un peu plus à un des partenaires qu'aux autres. Etant donné que nous sommes faits pour repérer les situations qui nous avantagent, les individus auront tendance à

repérer plus aisément les jeux à somme non nulle qui les avantagent plus que les autres. Ou plutôt, admettant que les individus ont des capacités globalement équivalentes, tout jeu à somme non nulle tendra à être repéré en premier par l'individu à qui il bénéficie le plus. Une fois que ce dernier a identifié cette possibilité, il doit en faire part à ses partenaires, qui devraient l'accepter. Cependant, il est possible que ses partenaires ne perçoivent pas immédiatement les avantages offerts par le jeu à somme non nulle. Dans ce cas, ils auront tendance à rejeter une intention qui leur aurait été bénéfique.

Il faut souligner que dans ces deux cas l'émetteur y perd également, car il n'arrive pas à faire faire (ou faire croire) ce qu'il veut au récepteur. Il y a donc des pressions de sélection, des deux côtés de la communication, pour trouver des mécanismes de vigilance épistémique fonctionnant de façon plus fine.

Calibrer la confiance

Une solution pour surmonter les limites de la vérification de cohérence est celle de la confiance : faire plus ou moins confiance à différentes personnes, et pour différents sujets. Les deux axes majeurs autour desquels s'organise l'évaluation sont ceux de la bienveillance et de la compétence (Fiske, Cuddy, & Glick, 2007). En fonction de nos liens avec une personne (apparenté, ami, allié, etc.), de nos relations passées avec elle, de son statut social, et de nombreux autres facteurs nous évaluons la confiance que nous pouvons lui accorder. Nous utilisons ensuite cette information pour évaluer les informations qui nous sont communiquées.

La confiance peut permettre de pallier les limites des mécanismes de filtrage basés sur l'incohérence. Voici comment. D'un côté, on prendra en compte l'éventuelle incohérence entre nos croyances préalables et ce qui nous est communiqué ; d'un autre côté, on prendra en compte la confiance que nous portons à la personne qui nous les communique. On peut ainsi parvenir à un équilibre. Ce mécanisme a cependant lui aussi des limites. Il peut-être long à mettre en place : il serait dangereux de se fier à une personne avec laquelle on n'a pas une assez longue histoire d'interactions nous indiquant qu'elle est généralement bienveillante et compétente. De plus, la tromperie est toujours possible : si la situation de l'autre a changé sans que nous ne le sachions, son intérêt à être bienveillant envers nous a

peut-être radicalement diminué. Il pourra alors abuser de notre confiance. Dans ce cas, nous accepterons une information qui pourrait nous être extrêmement dommageable. Il est alors possible pour le récepteur de baisser sa confiance en l'émetteur, voire de le punir par d'autres moyens, mais le mal sera déjà fait. De plus, les émetteurs doivent attendre d'avoir établi ces liens de confiance (ce qui peut prendre des années) avant de pouvoir essayer de communiquer des informations qui ne passeraient pas le filtre de vérification de cohérence. Ce système est donc loin d'être idéal. Il complémente bien les mécanismes basés sur la cohérence mais il partage leur principale limite : il est trop prudent, et empêche la communication de certaines informations. Après tout, il arrive que des gens pas toujours compétents ou pas toujours bienveillant aient à nous communiquer des informations qui entrent en conflit avec ce que nous pensions.

Plus généralement, on peut noter que quelques soient les limites d'un mécanisme de filtrage (vérification de cohérence, confiance, ou n'importe quel autre), ils devront toujours être plus aigus pour l'émetteur que pour le récepteur. Ou plutôt, ce mécanisme protège nécessairement les récepteurs : s'ils étaient perdants à la communication, ils arrêteraient simplement de recevoir. Donc, parmi les deux erreurs possibles – accepter trop d'informations dont certaines sont dommageables, et en refuser trop dont certaines seraient bénéfiques – c'est le second type qui prévaudra. Etant donné que les récepteurs rejettent nécessairement trop d'information à tout moment donné de l'évolution, des mécanismes plus sophistiqués ne peuvent faire qu'accroître la quantité d'informations qui seront acceptées.

1.3 L'argumentation comme moyen de surmonter les limitations des mécanismes de vigilance épistémique

Nous venons de voir que les mécanismes de vérification de cohérence et de confiance ont des limites. Ils empêchent les émetteurs de communiquer avec succès certaines informations, et ils privent par la même occasion les récepteurs de certaines informations qui pourraient leur être utiles. Des pressions de sélection se créent donc pour trouver une solution à ce problème.

Imaginons qu'un individu A demande à B de prendre une décision donnée. Cette décision ne correspond pas aux plans de B et il la rejette. A a alors intérêt à

essayer d'influencer B pour qu'il prenne tout de même cette décision. Il peut utiliser la communication en donnant des raisons pour lesquelles B devrait prendre cette décision. Pour comprendre comment une telle chose est possible il faut revenir à la façon dont ce qui est communiqué est évalué.

B évalue ce qui lui est communiqué par rapport aux plans et aux croyances qu'il a en tête. Ce sont (entre autres) les mécanismes qui régulent la compétition entre les intentions et qui gèrent la mise à jour des croyances qui sont utilisés pour évaluer si quelque chose de communiqué doit être accepté ou non. Ces mécanismes fonctionnent de façon très locale : leur objectif n'est pas de maintenir la cohérence globale du système cognitif (bien qu'ils y contribuent), mais d'empêcher d'une part que des actions incohérentes soient entreprises et d'autre part que des croyances qui ne correspondent plus à l'état du monde persistent dès lors que de nouvelles croyances sont obtenues. Ils ne cherchent pas de conflits entre des intentions ou des croyances, ils se contentent de les régler quand ils se produisent afin d'empêcher qu'ils n'aient des conséquences fâcheuses. Etant donné que ces mécanismes attendent que de tels conflits apparaissent et ne les recherchent pas activement, il est tout a fait possible pour les individus d'avoir des connaissances (et a fortiori des intentions) incohérentes, pour peu qu'elles ne soient pas activées au même moment : seules les intentions ou les croyances qui sont activées simultanément (ou en proche contiguïté temporelle) peuvent ainsi entrer en conflit. Ce n'est donc pas l'ensemble des intentions ou des croyances d'un individu qui vont déterminer s'il accepte ou rejette ce qui lui est communiqué, mais les intentions et les croyances qui sont le plus accessibles étant donné son état d'esprit et ce qui vient de lui être communiqué (même s'il n'était pas en train de penser à ce qui lui est communiqué, ce qui vient d'être communiqué attire nécessairement son attention sur son contenu).

L'individu A peut donc essayer de modifier les éléments qui seront pris en compte durant l'évaluation de ce qui va être communiqué. A devra trouver des éléments que B serait prêt à accepter (sur la base de la confiance ou parce qu'ils ne sont pas incohérents avec ses intentions ou croyances préalables) et qui augmenteront les chances qu'il accepte ce que A voulait communiquer en premier lieu. En d'autres termes, A doit trouver des raisons pour convaincre B. Le prochain chapitre se penchera sur la façon dont A peut réussir à trouver de telles raisons, et dont B peut à son tour les évaluer.

Dans un premier temps, B n'a pas besoin d'évaluer les raisons qui lui sont offertes en tant que raisons. Admettons que A veuille entraîner B à aller attaquer un autre individu (Jean). Cependant, B n'est pas très motivé (il a d'autres choses à faire), et il refuse d'accompagner A dans son attaque. A peut alors donner à B une raison pour le faire tout de même, telle que :

1 Jean a insulté ta mère
2 Allons attaquer Jean

Admettons par ailleurs que B savait que Jean avait insulté sa mère, mais qu'il n'y pensait plus à ce moment là. Lorsque A le lui rappelle, cette remémoration peut suffire à le mettre dans un état d'esprit tel qu'il accepte finalement d'aller attaquer Jean. Dans ce cas, B n'a pas eu à user de raisonnement lui-même. Il se peut très bien qu'il n'ait perçu le lien entre 1 et 2 que de façon intuitive : soit que 1 ait directement déclenché une inférence vers 2, soit que 1 ait augmenté les chances qu'il accepte 2 en modifiant son état d'esprit vis-à-vis de Jean. Il est également possible que B utilise le raisonnement pour évaluer la qualité de 1 en tant qu'argument, en tant que raison pour accepter 2. B serait encore plus clairement amené à raisonner si A lui avait dit 3 au lieu de 1-2 :

3 Allons attaquer Jean car il a insulté ta mère

En raisonnant sur ce que lui dit A, B accroît la probabilité d'être influencé par l'argument s'il est bon, et de ne pas l'être s'il est mauvais. De son côté, A a donc tout intérêt à utiliser des outils linguistiques (donc, car, si…alors, etc.) pour mettre l'accent sur ses raisonnements. Mais pour cela, encore faut-il être capable de trouver des raisons, et des bonnes. Nous allons voir dans le prochain chapitre comment un tel exploit peut être accompli.

2 Fonctionnement du raisonnement

Nous avons vu dans le chapitre précédent qu'à un moment de l'évolution de la communication il a pu devenir avantageux pour les individus de chercher des raisons : des éléments qui augmentent les chances qu'un autre individu accepte une information donnée (qu'on peut appeler une conclusion). Nous allons voir qu'une telle tâche implique d'être également capable d'évaluer des raisons, et c'est là pour nous la fonction primaire du raisonnement : évaluer des raisons. Cette capacité pourra également être utilisée par les récepteurs pour évaluer les arguments qui leurs sont présentés. On peut ouvrir ce chapitre sur un exemple banal. Pierre et Marie veulent aller au cinéma :

Pierre : prenons le métro
Marie : non, allons-y plutôt à pied

Pierre aimerait convaincre Marie. Il doit pour cela trouver d'autres éléments qui rendront son option plus cohérente avec les croyances ou les intentions de Marie, ou qui rendront l'option de Marie incohérente avec certaines autres de ses croyances ou intentions. Comment trouver ces raisons ? Pour comprendre comment s'y prendre, on peut s'intéresser au fonctionnement d'un mécanisme qui a lui aussi, parfois, besoin de chercher des éléments précis parmi de nombreux autres : la recherche visuelle.

2.1 Quelques éléments sur la recherche visuelle

Nous sommes souvent confrontés à des tâches de recherche visuelle : trouver ces clés que nous avons égarées, chercher un livre dans les rayonnages de notre bibliothèque, reconnaître un ami dans une foule. La façon dont la recherche s'effectue dépend des mécanismes d'attention visuelle. Parmi ces mécanismes, on peut distinguer deux grands types : ceux qui sont basés sur la prédiction, et ceux qui fonctionnent suivant la méthode 'générer et tester'. Les mécanismes basés sur la prédiction sont passifs : il s'agit simplement d'attribuer davantage de saillance à certain stimuli dès qu'ils sont perçus. C'est ce type de mécanisme qui est utilisé

lorsque notre attention s'oriente vers un objet très coloré, ou qui ne se trouve pas là où on l'attendait. Ainsi, il est possible de percevoir, aussi vite que possible, des stimuli très pertinents, de manière générale, dans notre environnement – des objets se dirigeant vers nous, des prédateurs. Ces mécanismes de prédiction peuvent être régulés en fonction du contexte : tel objet qui nous paraîtra très saillant dans un contexte pourra passer inaperçu dans un autre (un clown dans un cirque ou à un enterrement). On peut dire qu'il s'agit de prédiction car la tâche de ces mécanismes est, en quelque sorte, de faire un pari sur le type de stimulus qui sera pertinent dans un environnement donné. Les mécanismes utilisant la méthode 'générer et tester', par contre, sont actifs. Ils utilisent la mémoire de travail afin de maintenir une certaine cible active, et parcourent l'environnement en recherchant cet objet. C'est ce qui nous permet par exemple de parcourir une foule du regard en y cherchant quelqu'un. Dans cette stratégie, des possibilités sont générées (par exemple, les différents visages examinés) puis testées (ces visages sont évalués pour savoir s'ils correspondent bien à celui de notre ami).

Quelle stratégie est la plus efficace dépend tout d'abord de la possibilité même de faire une prédiction : s'il s'agit d'un nouveau domaine – au niveau évolutionniste et développemental – il n'est pas possible de savoir à l'avance quel stimulus sera pertinent dans un contexte donné. La réponse dépend également des coûts requis pour faire une prédiction : même si une prédiction est possible, elle peut demander des calculs complexes par exemple. Pour ce qui est de la stratégie générer et tester, les coûts augmentent lorsque l'espace à explorer augmente de taille et que les coûts des tests augmentent. Il est ainsi plus dur de trouver un objet parmi de nombreux autres (augmentation de la taille de l'espace à explorer) et surtout si les autres objets ressemblent à la cible (les tests deviennent plus coûteux pour pouvoir être sûr qu'il s'agit de la cible et non d'un distracteur – voir Li, 2002).

Ces deux types de mécanismes ont chacun des avantages propres. Si la recherche passive n'existait pas, nous pourrions nous faire facilement surprendre alors que nous sommes concentrés sur une tâche donnée : notre attention ne pourrait jamais être attirée par des facteurs bottom-up, et nous heurterions la moitié des obstacles sur notre chemin pour peu que nous soyons en train de penser à autre chose. Pour ce qui est des mécanismes 'générer et tester', ont conçoit mal comment des espèces un tant soit peu sophistiquées pourraient s'en passer. Elles seraient alors

complètement ‘à la merci’ de leur environnement, une mère ne pourrait même pas chercher ses petits.

Dans la plupart des cas, ces deux types de mécanismes seront utilisés en combinaison. Imaginons que vous vouliez clouer quelque chose, mais que vous n’ayez pas de marteau à disposition. Vous cherchez un objet qui pourrait remplir la même fonction. La solution basée sur les mécanismes de prédiction serait de spécifier exactement l’objet voulu, et d’en déduire ses caractéristiques visuelles. Il doit s’agir d’un objet facile à prendre en main, qui ne soit pas fragile, qui ne soit pas précieux, etc. Inférer de ces traits les caractéristiques visuelles que doit avoir l’objet est une tâche très difficile, mais si elle est possible, il suffit ensuite de laisser notre attention être guidée, de façon bottom-up, par les objets qui possèdent ces caractéristiques. Une solution beaucoup plus plausible est d’utiliser un mélange des deux stratégies. Plutôt que de spécifier à l’avance toutes les caractéristiques nécessaires, seules les plus naturelles sont retenues, comme la taille (il faut que l’objet ne soit ni trop petit ni trop grand). Tous les objets d’une taille donnée sont ainsi rendus plus pertinents et, grâce à des mécanismes bottom-up de prédiction, notre attention se portera successivement sur ces différents objets. Il est ensuite possible, une fois que notre attention s’est portée vers un de ces objets, de l’évaluer sur les dimensions qui n’ont pas été prises en compte jusqu’à présent – sa solidité, sa facilité de prise en main, etc.[4]

Un trait qui ressort de cette analyse est l’importance des concepts. Ce sont eux que nous utilisons pour faciliter la recherche : plus une recherche active correspond à un concept préétabli, plus elle sera facile. L’exemple de l’outil improvisé est justement difficile car il ne correspond pas à une catégorie que nous avons déjà utilisé, du moins pas régulièrement (voir Barsalou, 1983, sur les catégories ad hoc). A l’inverse, rechercher une personne connue est beaucoup plus aisé. Les concepts jouent également un rôle différentiel : plus les distracteurs appartiennent à des catégories éloignées de celle de la cible, et plus ils tendent à appartenir à une même catégorie, plus la recherche est facile (voir Li, 2002, pour revue). Ainsi il sera plus facile de retrouver une balle de tennis mise par inadvertance

[4] Cette seconde phase est rendue nécessaire par le fait que l’ordre dans lequel les facteurs bottom-up nous présentent les objets sera influencé par des facteurs non pertinents tels que la couleur, ce qui fait que les objets ne se présenteront pas à nous dans un ordre qui respectent tous les critères recherchés.

dans le tiroir des ustensiles de cuisine que si nous recherchions le couteau à huître. De plus, retrouver n'importe quel objet dans le placard qui nous sert de débarras depuis des années sera difficile car il n'est pas possible de créer une catégorie perceptuellement identifiable à laquelle appartiendraient les distracteurs.

2.2 Le cas du raisonnement

Lorsque nous devons trouver un argument pour défendre une position, nous sommes dans un cas qui partage plusieurs points communs avec le cas de la recherche visuelle. A la place d'objets du monde, il s'agit de représentations. On retrouve les différents types de recherche : nous pouvons par exemple avoir une idée très précise de l'argument que nous recherchons – ça sera le cas lorsque nous cherchons en mémoire un argument connu. Dans la majorité des cas cependant nous ne pouvons pas savoir à l'avance exactement ce qui constituera un bon argument. La recherche d'arguments s'apparente plus à la recherche d'un outil improvisé : nous avons une vague idée de ce qui conviendrait, mais pas une idée assez précise pour contraindre fortement la recherche. Comme dans le cas des outils improvisés cette limite est due au fait que les catégories préexistantes ne sont pas appropriées. De même que nous n'avons pas de catégorie pour 'objets pouvant remplacer un marteau pour enfoncer un clou', nous n'avons pas de catégorie (avant d'être confronté à la tâche) pour 'représentations pouvant servir d'argument pour convaincre une personne qui veut marcher que le métro est une meilleure solution'. De telles catégories peuvent s'acquérir, soit par expérience, soit par apprentissage explicite, mais la variété des situations argumentatives rencontrées dans la vie de tous les jours fait que dans la majorité des cas nous devrons créer une nouvelle catégorie 'sur-mesure'.

De la nécessité de mécanismes spécifiques

Avant de se pencher sur la question des mécanismes qui sont utilisés par le raisonnement pour résoudre ce problème, il est bon de souligner les raisons mêmes qui font que ce problème existe : pourquoi les mécanismes d'attention préexistants

ne sont-ils pas suffisants ? Dans le cas de la recherche visuelle, la recherche passive ordonne les objets présents dans l'environnement de façon à ce que ceux qui sont, en général, les plus pertinents soient ceux vers lesquels notre attention tende à se tourner spontanément. La recherche active, avec maintien en mémoire de travail, permet une plus grande flexibilité : elle permet d'adapter la recherche à des demandes ponctuelles qui diffèrent significativement de ce qui est généralement pertinent. Dans les deux cas, ces mécanismes sont nécessaires car les objets de l'environnement ne vont pas se présenter spontanément dans un ordre optimal pour l'individu. Au contraire, proies et prédateurs vont avoir tendance à essayer de se dissimuler par exemple. Si ce n'était pas le cas, il suffirait d'attendre que les mécanismes perceptifs enregistrent ce qui se présente à eux. Il faut donc des mécanismes attentionnels pour compenser, et faire que ce qui est réellement pertinent soit ce qui est détecté le plus rapidement, ce qui attire le plus l'attention. Ce problème semble cependant ne pas se poser au niveau des représentations. Nos représentations sont ordonnées, s'organisent de façon à ce que celle qui est la plus pertinente à un moment donné tende à être celle qui sera justement activée. C'est précisément ce que font les mécanismes attentionnels et autres mécanismes de régulation (le principe cognitif de pertinence Sperber & Wilson, 1995). Il faut souligner que les propriétés qui sont intéressantes dans ce contexte sont des propriétés des objets du monde : sont-ils dangereux, sont-ils comestibles, etc. La représentation d'un objet doit être pertinente à proportion de l'importance de l'objet lui-même. Bien entendu ces mécanismes ne font qu'approximer une répartition optimale des ressources, mais ce cas est tout de même radicalement différent de celui des objets du monde, qui eux ne s'ordonnent absolument pas d'une façon qui nous est naturellement avantageuse.

Même les mécanismes les plus flexibles – tels que la mémoire de travail – continuent de porter sur des propriétés des objets du monde. Lorsque nous recherchons un objet dans notre environnement, ce sont ses propriétés qui sont pertinentes, et le rôle de la mémoire de travail et de la recherche active ('générer et tester') est précisément de rendre pertinents les traits qui sont pertinents pour cette recherche spécifique.

Les problèmes commencent à se poser lorsque des propriétés des représentations sont en jeu et en particulier, dans le cas du raisonnement, des propriétés telles que 'être un bon argument'. Dans ce cas, il n'y a pas de raison pour que les catégories correspondent exactement à ce qui est désiré. Les représentations

acquièrent des propriétés qui sont pertinentes pour le raisonnement (la propriété d'être un bon argument par exemple) de par la façon dont elles sont organisées, mais elles ne sont pas organisées autour de ces propriétés. Là encore, on peut faire une analogie avec les objets du monde[5]. Les objets du monde acquièrent des propriétés qui sont pertinentes pour tel ou tel mécanisme cognitif de par la façon dont ils sont organisés. Ainsi le fait d'être un visage, qui est pertinent pour un mécanisme de détection de visage, est dû aux lois causales qui gouvernent le monde et qui ont fait que certains objets sont des visages et ont certaines propriétés. Les objets du monde ne sont cependant pas organisés autour de ces propriétés. Tous les visages ne se regroupent pas spontanément dans notre champ visuel. Le rôle des mécanismes cognitifs est précisément d'organiser les représentations des objets du monde en catégories pour pouvoir y réagir de façon appropriée. De même, le rôle du raisonnement est d'ordonner les représentations[6] de telle façon que celles qui ont les bonnes propriétés soient retenues lorsqu'elles sont recherchées, ou de sélectionner celles qui ont les bonnes propriétés lorsqu'elles se présentent à lui.

Inefficacité des mécanismes de prédiction dans le raisonnement

Ayant établi la nécessité de mécanismes spécifiques nous permettant de trouver de bons arguments, on peut maintenant se demander lesquels sont les plus appropriés : des mécanismes passifs de prédiction, bottom-up, ou des mécanismes actifs basés sur le 'générer et tester' utilisant la mémoire de travail ? Je vais commencer par suggérer que des mécanismes de prédiction, tels qu'on les trouve dans le système visuel par exemple, seraient néfastes dans le cas du raisonnement. Je me tournerai ensuite vers les avantages qu'ont dans ce cas les mécanismes basés sur la stratégie de 'générer et tester'.

Dans le cas du système visuel la pertinence bottom-up, la pertinence générale de certaines catégories, est importante car il est bénéfique que notre attention soit parfois dirigée vers ces objets. Par exemple, un animal ne disposant que de ces mécanismes pourra parfaitement réagir à la perception d'une proie ou d'un prédateur – objets qui auront attiré son attention. Par contre, un animal que ne posséderait que

[5] Ici entendu à l'exclusion des représentations.

[6] Ce qu'il doit faire au moyen de représentations de représentations, de métareprésentations.

des mécanismes top-down, s'il pourrait rechercher un objet souhaité dans l'environnement, ne serait pas averti lorsqu'un objet potentiellement très pertinent est perçu ailleurs que dans le champ de son attention focalisée. Or dans le cas du raisonnement cette capacité n'est justement pas nécessaire – en fait, elle serait même extrêmement encombrante et coûteuse. L'équivalent des mécanismes bottom-up dans le cas du raisonnement serait l'activation d'arguments dès qu'une représentation est entretenue par le système cognitif. Pour chaque représentation activement traitée, tout argument potentiellement pertinent serait également activé. Or trouver de tels arguments n'apporte le plus souvent aucune amélioration d'un point de vue épistémique ou pratique (voir l'introduction de la section concernant les résultats empiriques), mais demande nécessairement un certain coût : des ressources sont diverties pour une tâche qui n'est pas du tout pertinente de façon générale. Trouver des arguments n'est utile que dans certains contextes bien spécifiques – lorsque nous voulons convaincre quelqu'un – et ce n'est donc que dans ces circonstances qu'il est bon d'en chercher. Il est alors nécessaire d'avoir à sa disposition des mécanismes top-down permettant d'orienter la recherche de tels arguments, car des mécanismes bottom-up seraient quant à eux le plus souvent néfastes.

Avantages des mécanismes 'générer et tester'

Dans le cas du raisonnement, plusieurs éléments militent en faveur de l'approche 'générer et tester'. Tout d'abord, l'espace des hypothèses à explorer est structuré d'une façon qui peut être utile au raisonnement. De manière générale, il y aura une assez bonne corrélation entre le fait qu'une représentation soit pertinente vis-à-vis d'une autre représentation et le fait qu'il s'agisse d'un bon argument potentiel. Cette observation est particulièrement vraie si on se place à l'échelle de l'ensemble de nos croyances par exemple : seul un nombre infime de nos croyances est rendu pertinent par l'activation de l'une d'entre elles, et il sera rare qu'une croyance qui n'a pas été rendue ainsi pertinente puisse en fait être un bon argument. Par contre, parmi les croyances rendues pertinentes, un nombre non négligeable ne sera pas approprié en tant qu'argument. On peut dire qu'une recherche orientée par des critères de pertinence généraux donnera beaucoup de faux positifs (une représentation jugée pertinente qui ne constitue pas en fait un bon argument) mais

peu de faux négatifs (une représentation qui n'est pas jugée comme étant pertinente alors qu'en fait elle aurait fait un bon argument). De plus, le coût des tests est très limité dans le cas du raisonnement. Pour comprendre pourquoi, il faut détailler la façon dont ils sont conduits.

Ici, l'analogie pertinente sera avec l'utilisation d'un outil pour savoir s'il dessert sa fonction correctement. Lorsque nous recherchons un outil improvisé, il est parfois trop dur de prédire son efficacité : il faut alors l'essayer pour savoir s'il convient en effet. Dans le cas du raisonnement, lorsqu'une représentation est considérée comme argument potentiel, il est de même possible de 'l'essayer'. Dans ce cas, nous utiliserons nos mécanismes métareprésentationnels afin de tester l'efficacité de l'argument envisagé. Par exemple, si nous envisageons d'argumenter en présentant une conséquence positive de l'option que nous défendons, il est possible d'activer la représentation correspondant à cette position et de voir si en effet elle active la conséquence souhaitée[7]. Ainsi, Pierre pourrait s'assurer que prendre le métro entraîne une arrivée rapide à destination.

Dans ce cas les coûts sont assez faibles. Tout d'abord, il ne s'agit que de coûts computationnels : nous n'allons pas nous faire dévorer s'il s'avère que la représentation testée n'est pas un bon argument. Par ailleurs, il ne s'agit pas d'une recherche engageant des calculs complexes : la partie la plus ardue (déterminer si une représentation en entraîne bien une autre) n'a pas besoin d'être calculée par des mécanismes indépendants qui seraient propres au raisonnement, elle est simplement simulée (de la même manière que le coût computationnel d'essayer un outil est moindre que celui d'essayer de calculer s'il remplira adéquatement sa fonction)[8].

Résumé sur le fonctionnement du raisonnement

[7] Ces deux représentations devant toute fois être maintenues sous forme de métareprésentation pour ne pas confondre cette activation avec une activation par un autre moyen, plus 'naturel' (non métareprésentationnel).

[8] Les mécanismes métareprésentationnels qui inhibent les inférences qui devraient normalement suivre de l'activation de ces représentations dépensent de l'énergie, mais cela reste un coût limité.

Les deux sections qui précèdent ont amené des éléments démontrant la supériorité de l'approche top-down 'générer et tester' dans le cas de la recherche de raisons – dans le cas du raisonnement. Une conséquence de cette stratégie est que lorsqu'on recherche une raison, on examinera différents candidats dans l'ordre dans lequel ils sont les plus pertinents. Le principal déterminant de cette pertinence sera la conclusion que nous voulons communiquer, mais d'autres facteurs contextuels – ce dont nous étions en train de parler, ce que nous avons à l'esprit en ce moment, ce que nous savons de notre interlocuteur, etc. – vont également influencer le degré de pertinence de différentes représentations. De plus, on peut envisager que certaines catégories d'arguments soient utilisées. Il est ainsi possible que nous ayons recours à des catégories telle que 'conséquence positive de la conclusion', ou au contraire 'conséquence négative de la position de l'interlocuteur', ou encore à d'autres catégories plus sophistiquées. Malgré l'aide que pourraient apporter ces catégories, les mécanismes de pertinence qui guident la recherche ne l'orientent pas spécifiquement vers de bonnes raisons, il est donc nécessaire d'évaluer les candidats afin de savoir s'ils constituent effectivement de bonnes raisons ou non. Pour cela, il est nécessaire d'avoir recours à des mécanismes métareprésentationnels.

Lorsqu'une représentation est examinée pour savoir si elle constitue une raison appropriée, il faut que nous nous la représentions en tant que raison (que nous la métareprésentions, donc). L'appareillage métareprésentationnel permet ensuite d'évaluer le lien entre la conclusion et la raison. Si par exemple la raison est censée entraîner la conclusion, l'activation de cette représentation devrait entraîner l'activation de la conclusion (par le biais d'une inférence intuitive[9]). Si le lien recherché est avéré, s'il est jugé suffisant, alors il est possible d'utiliser la raison examinée comme argument.

Ce mécanisme peut être utilisé également lorsqu'il s'agit d'évaluer un argument qui nous est adressé. Dans ce cas, la personne qui examine l'argument n'a pas besoin de procéder à une recherche de raisons : elles lui sont fournies par l'émetteur. Une fois qu'un énoncé est évalué comme raison, s'il est jugé comme étant une bonne raison, il augmente les chances que la conclusion qu'il soutien soit

[9] En tout cas si la raison est une croyance intuitive. S'il s'agit d'une croyance réflexive, il est possible que cette inférence soit elle-même réflexive (voir Sperber, 1997).

acceptée. J'avais mentionné plus haut la possibilité qu'un énoncé augmente les chances qu'un autre soit accepté sans pour autant qu'il ne soit considéré comme une raison. Cette possibilité permet qu'un énoncé puisse remplir ces deux rôles. Pour reprendre l'exemple utilisé alors, si A dit à B :

1 Jean a insulté ta mère
2 Allons attaquer Jean

1 peut augmenter les chances que 2 soit accepté sans l'aide du raisonnement, mais il aura encore plus de chances de le faire s'il est considéré en tant que raison. En effet, le fait d'évaluer 1 en tant que raison pour accepter 2 focalise l'attention sur les liens inférentiels éventuels entre 1 et 2. Il y a donc moins de chance que B ne passe à côté du lien auquel A voulait avoir recours, c'est-à-dire le lien entre le fait que quelqu'un ait insulté notre mère et la motivation pour lui vouloir du mal, s'il considère 1 comme une raison. B pourrait le faire spontanément, mais il y a beaucoup plus de chances qu'il le fasse si le lien est explicitement pointé par A en disant par exemple

3 Allons attaquer Jean car il a insulté ta mère

Cette discussion montre également que les effets du raisonnement sont de nature épistémique. Son effet direct est de rendre un verdict sur l'acceptabilité d'une représentation en tant que raison pour accepter (ou rejeter) une conclusion. C'est là l'effet unique qui est recherché par l'émetteur pour qui les effets devraient normalement en rester là : étant donné que la représentation qui servira de raison était déjà présente dans son système cognitif, le fait de la considérer en tant que raison ne devrait pas, dans des circonstances normales, avoir d'influence sur le statut épistémique de la conclusion. Le fait de trouver des arguments pour une position que nous tenons déjà ne devrait pas, en principe, avoir d'effet sur cette position même[10]. Par contre, pour le récepteur, l'effet indirect de cette évaluation, *doit être* d'augmenter les chances qu'une conclusion donnée soit acceptée (ou rejetée) (voir Mercier & Sperber, in press).

[10] Nous verrons qu'il y a de nombreuses exceptions à ceci, en particulier lorsque le raisonnement est utilisé en dehors de son contexte normal, voir la section 7.7.

Au début de la seconde partie les prédictions que fait la théorie argumentative sur le fonctionnement du raisonnement seront reprises afin de pouvoir les comparer à la littérature empirique, mais avant ceci on peut comparer les résultats de cette partie théorique aux autres théories à processus duel qui ont été proposées dans le domaine du raisonnement.

3 Comparaison avec les autres théories à processus duel

L'objectif de ce chapitre est de faire une revue des autres théories à processus duel dans le domaine du raisonnement, afin de pouvoir comparer leurs prédictions à celles de la théorie argumentative. Cette revue permettra également de découvrir certaines des expériences qui sont généralement avancées comme soutien pour les théories à processus duel. Enfin, il s'agira également de l'occasion d'explorer certaines conséquences de la théorie argumentative, en particulier pour ce qui est de la place du raisonnement dans l'organisation générale de l'esprit et de la façon de concevoir les hypothèses évolutionnistes.

3.1 La théorie d'Evans

Les origines

Comme tant de choses dans la psychologie du raisonnement, les théories à processus duel furent pour la première fois suggérées par Peter Wason, dans un article cosigné par leur futur grand défenseur, Jonathan Evans (Wason & Evans, 1975). Sur quelle base les auteurs suggèrent-ils l'existence de deux types de processus distincts ? Il faut pour le comprendre expliquer rapidement la tâche qu'ils ont utilisée, tâche qui sera de toute façon, de par la place centrale qu'elle joue en psychologie du raisonnement, récurrente dans cette thèse. Il s'agit de la fameuse tâche de sélection (aussi connue sous le nom de 'tâche de sélection à quatre cartes', où, d'après son créateur, de 'tâche de sélection de Wason' (Wason, 1966). Dans sa forme la plus classique, cette tâche demande aux participants de vérifier si une règle est vraie de quatre cartes disposées face à eux. Chacune de ces cartes porte une lettre sur une face et un chiffre sur l'autre, une seule des deux faces étant visible des participants. La règle originale, abstraite, était la suivante : « Si une carte a une voyelle d'un côté, alors elle a un nombre pair de l'autre ». La tâche des participants est donc de vérifier que cette règle est vraie en ne retournant que les cartes nécessaires, en retournant le moins de cartes possibles. Les faces visibles montrent

les éléments suivants : 4, E, K et 7. La bonne réponse consiste à retourner les cartes E et 7, mais seuls 10 % des participants tendent à la trouver, les erreurs les plus courantes consistant à oublier le 7 ou à ajouter le 4. Retourner la carte portant le 7 est nécessaire car si elle porte une voyelle sur son autre face, alors la règle est falsifiée. La carte portant le 4 ne l'est pas car même si elle portait une consonne sur son autre face, la règle n'en serait pas infirmée pour autant.

Parmi les premières observations à être faite sur cette tâche on trouve le phénomène d'appariement (ou 'matching', Evans & Lynch, 1973). Lorsque l'énoncé comporte des négations (par exemple « Si une carte a un A d'un côté, alors elle n'a pas un 5 de l'autre »), les réponses données tendent à correspondre aux éléments mentionnés dans la règle – d'où le terme d'appariement. Or, dans le cas de la règle qui vient d'être donnée, cette réponse (A et 5) est en fait la bonne. Il est cependant douteux que le simple fait de mentionner une négation permette aux participants de soudainement mieux comprendre la tâche. Il semble plus probable que dans tous les cas les participants tendent à apparier leurs réponses aux éléments mentionnés dans la règle, sans réellement comprendre que cette opération les mène, parfois, à la bonne réponse[11]. Afin de tester cette hypothèse, Wason et Evans ont confronté les participants à ces deux conditions – une dans laquelle l'appariement mène à la bonne réponse et une à la mauvaise – mais en leur demandant de justifier leurs réponses, carte par carte.

Ils observèrent que les participants avaient d'avantage tendance à produire des justifications correctes (en termes de falsification) dans le cas ou le conséquent était nié – donc le cas où leur réponse était en effet correcte. Une première lecture des résultats semble donc indiquer une réelle compréhension de la part des participants. Wason et Evans rejettent cependant cette interprétation pour la raison suivante : de ces mêmes participants, qui semblent avoir compris la tâche, aucun ne fût ensuite capable de donner la bonne réponse avec la règle classique, n'impliquant pas de négation. Il serait bien étrange que des participants qui aient parfaitement compris la tâche se retrouvent soudainement totalement dépourvus simplement parce que la règle ne mentionne pas de négation. Les auteurs expliquent le recours à des

[11] Ces explications en termes d'appariement ont depuis été remplacées par des explications plus sophistiquées – en termes de pertinence par exemple (Sperber, Cara, & Girotto, 1995), mais cela ne change rien au point exposé ici car dans tous les cas les processus correspondant à l'appariement sont différents de ceux menant à une compréhension réelle de la tâche.

explications en termes de falsification par une forme d'amorçage par la négation : le fait d'avoir un élément nié dans la règle rendrait plus disponibles des explications impliquant elles aussi des termes négatifs (« la règle est fausse » « pour être sûr qu'il n'y a pas 5 », etc.).

Sur cette base, Wason et Evans proposent la première théorie à processus duel du raisonnement : un type de processus serait responsable du phénomène d'appariement, et un autre des processus de justification (qui ne sont ici guère plus que des rationalisations). Cette théorie remet alors en cause un résultat précédemment obtenu par Wason (Wason, 1969). Il avait observé dans cette expérience que s'il donnait la solution correcte aux participants, ceux-ci étaient alors parfaitement capables de l'expliquer. Ils semblaient donc avoir bien compris la tâche. Mais à la lumière de ces nouveaux résultats, une autre interprétation est possible : les participants ne feraient que rationaliser la réponse donnée, réponse qu'ils auraient acceptée uniquement car l'expérimentateur venait de leur dire qu'elle était la bonne. Afin de tester cette hypothèse, Evans et Wason conduisirent une nouvelle expérience (Evans & Wason, 1976). Dans celle-ci les participants devaient également justifier une réponse qui leur était donnée comme bonne – mais cette fois cette réponse pouvait très bien ne pas être, en fait, correcte. Et, comme prédit, les participants furent capables, dans tous les cas, d'expliquer pourquoi il s'agissait de la réponse 'correcte', confirmant qu'il s'agissait bien de rationalisations et non d'une réelle compréhension de la résolution du problème.

On trouve dans l'article original de Wason et Evans plusieurs éléments qui seront appelés à jouer un rôle important dans le développement des théories à processus duel – et même, pourrait-on dire, de la psychologie plus généralement. Le premier est bien entendu l'idée même de théorie à processus duel. Le second est l'idée de rationalisation. A cette époque plusieurs expériences de psychologie sociale et de résolution de problème pointent vers l'idée que les réponses explicites des participants ne sont souvent que des justifications post-hoc de leurs décisions et ne jouent dans celles-ci aucun rôle causal (voir Nisbett & Wilson, 1977). Bien que les résultats de Wason et Evans ne soient pas cités par Nisbett et Wilson, ils vont clairement dans la même direction, une direction qui sera très influente (on peut même sûrement dire de plus en plus influente) en psychologie. Enfin, les auteurs entrevoient déjà la possibilité que les relations entre les deux types de processus soient plus compliqués qu'il n'y paraît :

> In its strongest form the dual process hypothesis, that response determines conscious thought, may be an oversimplification. A weaker (but more plausible) assumption is that there is a dialectical relation between them: A process of rapid continuous feedback between tendencies to respond and consciousness rather than two temporally distinct phases. (Wason & Evans, 1975, p.150)

Malgré ses aspects quasi 'prophétiques' (toujours plus faciles à discerner post hoc), cet article sera quasiment ignoré jusqu'au milieu des années 90, moment où les théories à processus duel font leur grand retour sur la scène, à travers un article de Steven Sloman et un livre de Jonathan Evans et David Over.

La théorie d'Evans et Over, circa 1996

Une place prépondérante sera accordée ici à la théorie développée par Evans et certains de ses collègues, en particulier Over. Il s'agit de la théorie la plus ancienne – comme nous venons de le voir, elle a ses sources dans les années 70 – mais c'est également la plus sophistiquée et celle qui tente de rendre compte assez précisément du plus grand nombre de données en psychologie du raisonnement. L'ordre chronologique ne sera pas scrupuleusement suivi, et les premières versions de la théorie ne seront que survolées ; la version la plus récente, au contraire, sera étudiée dans le détail. Les autres théories – celles de Sloman, et de Stanovich – seront également vues plus tard, même si leurs premières apparitions sont antérieures à la version de la thèse d'Evans présentée ici.

On retrouve des allusions à une théorie à processus duel à intervalles réguliers de la carrière d'Evans (voir en particulier Evans, 1984), mais ce n'est qu'en 1996, dans un ouvrage écrit avec David Over, qu'il exposera dans le détail sa théorie pour la première fois. Les auteurs se détournent de la distinction entre processus orientant vers la réponse et simples processus de rationalisation qui avait été introduite par Wason et Evans en 1975, et prennent cette fois pour point de départ, ou pour exemple principal, le biais de croyance. Cet effet apparaît lorsque, dans un raisonnement syllogistique par exemple, la logique et les croyances préalables dictent

des réponses contradictoires. Le biais de croyance est alors l'effet résiduel de la croyance préalable sur la réponse des participants. Afin d'illustrer cet effet on peut se contenter de reprendre des problèmes tirés de l'article l'ayant le premier mis en lumière (Evans, Barston, & Pollard, 1983). Voici les quatre catégories d'arguments :

Argument valide, conclusion crédible

Aucun chien policier n'est vicieux
Certains chiens très entraînés sont vicieux
Donc, certains chiens très entraînés ne sont pas des chiens policiers

Argument valide, conclusion non crédible

Aucune chose nutritive n'est peu chère
Certaines tablettes de vitamines sont peu chères
Donc, certaines tablettes de vitamines ne sont pas nutritives

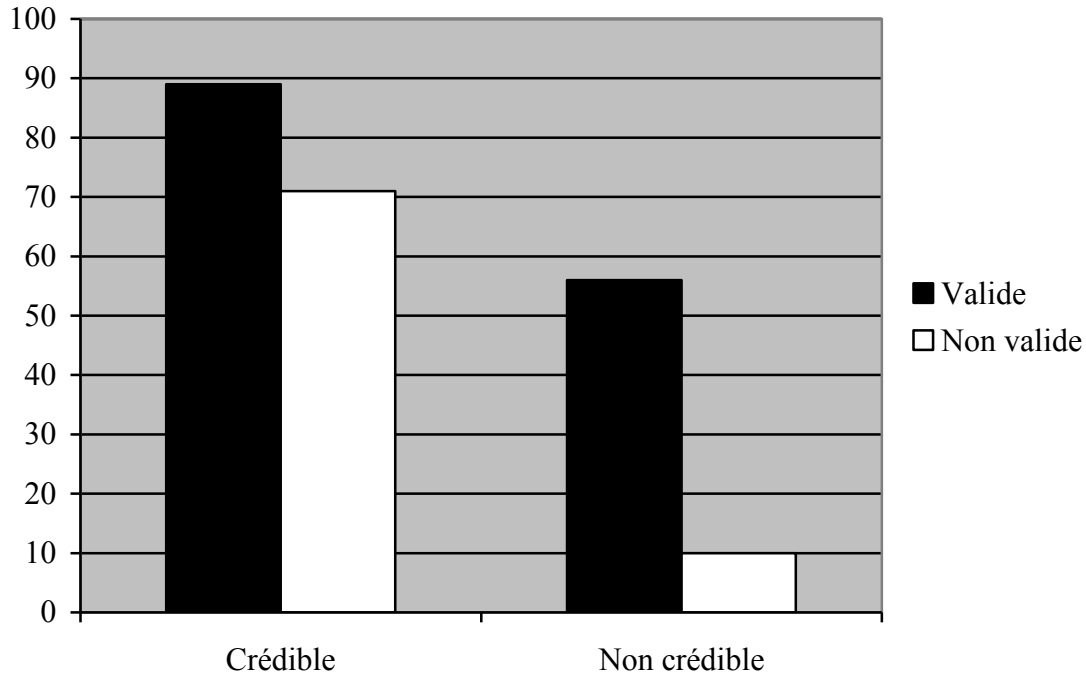
Argument invalide, conclusion crédible

Aucune drogue n'est peu chère
Certaines cigarettes sont peu chères
Donc, certaines drogues ne sont pas des cigarettes

Argument invalide, conclusion non crédible

Aucun millionnaire n'a un travail difficile
Certains gens riches ont un travail difficile
Donc, certains millionnaires ne sont pas des gens riches

Et voici les résultats obtenus :



Est ici représenté le taux d'acceptation de la conclusion (en pourcentage), en fonction de sa crédibilité et sa validité (adapté de Evans et al., 1983) (figure tirée de Noveck, Van der Henst, Rossi, & Mercier, 2007).

On peut voir sur ce graphique (et les analyses statistiques le confirment) deux effets différents : d'une part les conclusions valides tendent à être plus acceptée que les conclusions invalides. Mais au-delà de cet effet de la logique les croyances préalables jouent également un rôle, et les arguments ayant des conclusions croyables tendent eux aussi à être mieux acceptés. Il s'agit là pour Evans et Over d'un élément militant en faveur de deux types de processus : un premier type qui serait influencé (malgré les instructions) par les croyances préalables, et un second qui serait lui capable de prendre en compte la validité des arguments. Ces deux types de processus sont nommés *heuristique* et *analytique*. Heuristique car il s'agit de mécanismes qui, bien que généralement efficaces, sont imparfaits et amènent à des erreurs régulières ; analytique car ce second système serait capable au contraire d'analyser, à l'aide de règles, les éléments d'un énoncé ou d'une décision à prendre afin de parvenir à une réponse correcte.

Il est intéressant de noter une sorte de renversement dans la vision de ce second type de processus. Alors que dans l'article original de Wason et Evans il

s'agissait quasi-entièrement d'un mécanisme responsable de la rationalisation – et donc déprécié, inutile voire nuisible – il s'agit ici au contraire du mécanisme même nous permettant de suivre les règles de la logique. Ces deux interprétations n'ont jamais été réellement contradictoires, car de la place était laissée pour un rôle plus glorieux dans l'article original d'un côté, et la possibilité de rationalisation est conservée dans les positions plus récentes (voir en particulier Evans, 2007, chapitre 7), mais nous assistons tout de même là à un changement assez important.

Un changement d'emphase similaire s'opère du côté des processus heuristiques. Ces processus sont généralement jugés responsables des erreurs de raisonnement : c'est le cas dans la tâche de sélection de Wason (dans laquelle les mécanismes analytique ne font, au pire, qu'échouer à les corriger), c'est le cas dans le biais de croyance, et ça l'est dans l'immense majorité de la littérature sur les heuristiques et les biais (Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982). Mais Evans et Over sont bien forcés de reconnaître une certaine efficacité à ces processus : après tout, ce sont eux qui guident la majorité de nos actions, et nous ne faisons pas que des erreurs. Au contraire, la plupart de nos actions (ou de nos actes de pensée plus généralement) accomplissent leurs objectifs d'une façon remarquablement efficace, étant donné l'immense complexité de certaines des tâches qu'ils affrontent (qu'il s'agisse de la simple coordination des mouvements, de la reconnaissance de divers objets ou bien de la production et compréhension d'énoncés). Cette reconnaissance de l'efficacité des mécanismes heuristiques n'empêche pas que dans la majorité des problèmes de raisonnement et de prise de décision, ce sont bien les processus analytiques qui donnent la réponse 'normative', celle qui est jugée comme étant correcte[12].

Bien qu'on ressente souvent une certaine préférence pour les processus analytiques chez Evans et Over[13], il n'en reste pas moins qu'ils reconnaissent aux deux types de processus une certaine rationalité. Cependant, processus analytiques et heuristiques donnent des réponses contradictoires dans de nombreuses tâches, et deux réponses contradictoires ne peuvent guère être jugées rationnelles en même temps, tout au moins si on utilise un seul et même critère de rationalité. Pour

[12] Même s'il y a sûrement un fort biais pour utiliser précisément des tâches dans lesquelles les mécanismes heuristiques donnent de mauvais résultats – voir la conclusion.

[13] Préférence qui sera encore plus marquée chez Stanovich, chez qui elle prend même une certaine dimension morale.

résoudre ce problème, Evans et Over introduisent deux critères de rationalité différents, chacun étant censé s'appliquer à un type de processus. Aux processus heuristiques, la Rationalité1 : « Penser, parler, raisonner, prendre une décision ou agir d'une façon qui est généralement fiable et efficace pour achever ses propres objectifs. » Et aux processus analytiques la Rationalité2 : « Penser, parler, raisonner, prendre une décision ou agir lorsqu'on a une raison pour ce que l'on fait sanctionnée par une théorie normative. » (Evans & Over, 1996, p.8).

Si on voit déjà émerger dans ses grandes lignes une théorie à processus duel, plusieurs aspects en restent encore vagues. Evans et Over rapprochent leur théorie d'autres théories proposées dans des domaines connexes (apprentissage, prise de décision, psychologie sociale), et reprennent à leur compte un certain nombre de caractérisations que des théories font des deux types de processus. Ils s'appuient en particulier sur les recherches dans le domaine de l'apprentissage implicite (Berry & Dienes, 1993; Reber, 1993). Ces travaux reposent en grande partie sur la distinction entre phénomènes conscients et inconscients[14] : ils se préoccupent principalement de certains types d'apprentissage dont les participants ne sont pas conscients, alors qu'ils ont de claires manifestations comportementales (amélioration des performances). Cette distinction entre processus conscients et inconscients, cependant, ne nous aide guère à mieux cerner le fonctionnement des deux types de processus. Les autres caractéristiques attribuées aux deux types de processus – évolutionnairement ancien ou récent, rapide ou lent, coûteux ou non, etc. – le seront de même sur la base de simples observations et suppositions qui semblent justifier ces généralisations. Il ne s'agit pas là de caractéristiques qui pourraient jouer un rôle fonctionnel, mais plus de corrélations, qui pourraient tout aussi bien être accidentelles et pour lesquelles, en tout cas, aucune explication n'est réellement fournie.

Un autre point sur lequel la théorie reste, à ce stade tout au moins, plutôt vague est celui des interactions entre les deux types de processus. Evans et Over se contentent en effet de critiquer un hypothétique modèle purement séquentiel, dans lequel les processus heuristiques précéderaient toujours, dans leur fonctionnement, les processus analytiques en admettant la possibilité d'effets en retour et autres interactions. Ils ne dépassent donc guère la position de l'article original de 1975.

[14] Qu'Evans et Over appellent 'tacites' – pour se différencier de notions freudiennes ?

Nous verrons qu'un des avantages de la théorie argumentative est justement d'expliquer pourquoi certains des traits prêtés à chacun des deux types de processus correllent en effet, ainsi que de faire des hypothèses plus précises sur la façon dont les deux systèmes interagissent.

Version la plus récente de la théorie d'Evans

Si la façon dont sont caractérisés les deux types de processus conservera ce même aspect de corrélation entre traits sans réelle explication (voir par exemple Evans, 2007, pp.14-15), les règles qui régissent le fonctionnement du système analytique, ainsi que les interactions entre les systèmes seront, elles, élaborées dans les versions plus récentes du modèle[15]. Evans n'a que peu de choses à dire sur le fonctionnement du système 1, et il est dur de le lui reprocher : il reconnaît (suivant Stanovich) qu'il s'agit en fait d'un ensemble de mécanismes ayant des fonctionnements différents, et il n'est donc guère surprenant de ne pouvoir dégager de règles communes, au-delà de vagues caractérisations – rapide, peu coûteux, etc. Il ne faut pas oublier non plus qu'Evans est un psychologue du raisonnement, et il est donc normal qu'il s'attarde principalement sur ce qui constitue (pour la théorie argumentative du moins) le 'vrai' raisonnement, à savoir les processus analytiques. Il apporte ici quatre avancées par rapport aux versions précédentes. Premièrement, il utilise le cadre de la 'pensée hypothétique' : il défend l'idée que les processus analytiques correspondent, par nature, à un mode de pensé hypothétique. Deuxièmement, il caractérise ce mode de pensée par trois règles : le principe de singularité, le principe de pertinence et le principe de satisficing. Troisièmement, il tente de réduire à deux biais fondamentaux les erreurs dont les systèmes analytique et heuristique peuvent se rendre coupable. Enfin, il formule des hypothèses plus précises sur la façon dont les deux types de processus interagissent. Je me contenterai ici de reprendre les grandes lignes des explications fournies par Evans et n'exposerai pas les données empiriques qu'il invoque pour les défendre, données dont la majorité sera passée en revue dans la seconde partie de cette thèse. Le court exposé de chacun de ces points sera suivi de la position que prend la théorie argumentative sur ce

[15] Les explications qui suivent sont principalement basées sur Evans (2006, 2007).

même point, ce qui démontrera, derrière un vocabulaire parfois différent, d'assez grandes similarités.

Pensée hypothétique

A la suite de ses travaux sur les conditionnels, Evans fait l'hypothèse que les mécanismes analytiques desservent la *pensée hypothétique*, c'est-à-dire « la pensée qui requiert l'imagination d'états possibles du monde. » (Evans, 2007, p.17). Ce type de pensée utiliserait des « modèles mentaux épistémiques », qui « représentent des états de croyance et de connaissance » (Evans, 2006). Il prend grand soin de distinguer ces modèles mentaux épistémiques des autres 'modèles mentaux', ceux qui sont éponymes à la théorie (concurrente) de Johnson-Laird. Pour Evans, l'interprétation que fait Johnson-Laird des modèles mentaux est ambiguë entre des modèles hypothétiques (proches des siens) et des modèles 'sémantiques', représentant directement certains aspects du monde. Il s'agit selon lui d'une erreur, une ligne devant être tracée justement entre ces deux interprétations. Evans distingue aussi ses modèles mentaux épistémiques de la simulation mentale. Pour lui, si la simulation mentale peut-être utile pour effectuer certains traitements portant sur ces modèles, elle n'est nullement nécessaire (Evans, 2007, p.17 et Evans, 2006).

Il me semble que ces distinctions peuvent être clarifiées en utilisant le vocabulaire de 'représentation' et 'métareprésentation'. Les représentations seraient les modèles mentaux 'sémantiques', représentant directement des aspects du monde. Qu'en est-il des modèles mentaux épistémiques ? Pour Evans, ils « encodent des attitudes propositionnelles vis-à-vis des états du monde » (Evans, 2006). Les attitudes propositionnelles sont des attitudes telles que « je pense que X », « je doute que X », « je suis sûr de X », etc. Or de telles attitudes propositionnelles sont métareprésentationnelles, puisqu'il s'agit de porter un jugement sur des représentations. Il semble donc qu'on puisse identifier la pensée hypothétique avec un certain type de métareprésentations.

Le fait d'exprimer la distinction posée par Evans en termes de métareprésentations permet d'en souligner une limitation : l'absence de distinction, au sein de ces mécanismes de pensée hypothétique, entre différents types de mécanismes. On pense en effet qu'il existe plusieurs types de mécanismes

métareprésentationnels, servant des fonctions différentes – pour la mentalisation, le langage et, justement, le raisonnement (voir par exemple Sperber, 2000). Il n'est donc pas clair si Evans souhaite limiter la catégorie des mécanismes hypothétiques à ceux traitant purement de raisonnement, ou s'il préfère l'étendre à tous les mécanismes métareprésentationnels. Par exemple, il est raisonnable de penser que la possibilité d'entretenir des attitudes propositionnelles a évoluée tout d'abord en visant les représentations d'autres individus (« Jean pense que », « Emmanuel doute que », etc.) (Byrne & Whiten, 1988; Sperber, 2000; Whiten & Byrne, 1997). On devrait alors inclure les mécanismes de mentalisation dans ceux de la pensée hypothétique telle qu'elle est décrite par Evans – une ouverture à laquelle il ne s'opposerait peut-être pas, étant donné les rapprochements qu'il tente de faire avec les théories à processus duel présentes en psychologie sociale. Il n'en reste pas moins que les résultats auxquels Evans applique son modèle sont entièrement tirés de la psychologie du raisonnement et de la prise de décision – on peut donc également défendre une position selon laquelle la pensée hypothétique serait plutôt réservée à ces domaines. Quoiqu'il en soit, si en effet différents mécanismes métareprésentationnels ont des histoires évolutives, développementales et un fonctionnement très différent, ne pas les différencier du tout risque d'occasionner une certaine confusion.

Finalement, notons un point de convergence : la distinction entre pensée hypothétique et simulation mentale. Le problème est ici que le terme de 'simulation mentale' peut recouvrir des mécanismes psychologiques qui n'ont que peu de choses en commun. Dans le cadre de la théorie argumentative, l'évaluation des arguments potentiels – pour l'émetteur – ou des arguments soumis – pour le récepteur – se fait par un mécanisme qu'on peut qualifier de simulation : il s'agit d'utiliser nos mécanismes métareprésentationnels afin de tester un lien entre deux représentations. Cette forme de simulation peut être, et est généralement, totalement inconsciente : elle a lieu lorsque nous réfléchissons pour trouver un argument, par exemple, sans nécessairement que nous ne nous entendions essayer, dans notre tête, différents arguments. Pour reprendre l'analogie présentée plus haut avec la recherche visuelle, si nous cherchons un outil improvisé en parcourant une pièce du regard, nous pouvons 'simuler' l'efficacité de différents objets de façon inconsciente, sans nous représenter consciemment en train d'essayer d'utiliser l'objet pour accomplir la tâche souhaitée. En ce sens, la simulation mentale accompagnera la grande majorité des

raisonnements, dans la mesure où ne pas s'en servir revient à prendre la première solution suggérée par les mécanismes de pertinence sans même s'assurer un tant soit peu qu'elle convienne. Mais, à nouveau, il est important de distinguer ce type de simulation de simulations conscientes du type 'se mettre à la place de quelqu'un pour essayer de savoir ce qu'il ferait' ou 'représenter une scène absente pour y chercher un objet'. Si ce type de mécanisme peut en effet jouer un rôle dans le raisonnement, Evans a sûrement raison de penser qu'il n'est nullement nécessaire. Et étant donné que c'est ce type d'acception qui est la plus courante, et qui est probablement celle visée par Evans lorsqu'il parle de simulation mentale, il n'y a pas ici de désaccord entre sa théorie et la théorie argumentative.

Pour conclure sur la pensée hypothétique, on peut dire que dans la mesure où Evans accepterait une réinterprétation de sa théorie en termes de métareprésentations, on observe une convergence assez frappante avec la position de la théorie argumentative. Se poursuit-elle lorsqu'il s'agit des principes censés régir le fonctionnement de cette pensée hypothétique ?

Principe de singularité

Evans fournit trois grands principes s'appliquant à la pensée hypothétique – et donc au système analytique. Le premier est le principe de singularité, selon lequel « lorsque nous pensons hypothétiquement nous ne considérons qu'une possibilité ou modèle mental à la fois » (Evans, 2007, p.17). Cette limitation aurait selon lui deux origines : d'une part les limitations de la mémoire à court terme, essentielle au fonctionnement des processus analytiques, et d'autre part leur nature fondamentalement sérielle. On peut accorder ces deux éléments, mais ils n'offrent, tels quels, qu'une explication proximale. En biologie, on parle d'explications 'proximales' et 'ultimes' pour différencier d'un côté les explications en termes de mécanismes présents à un moment donné et d'un autre celles en termes de causes évolutives, contraintes et pressions de sélection. On peut par exemple expliquer de deux façons la relative pauvreté de l'odorat humain : au niveau proximal, on peut constater que nos organes olfactifs sont sous-développés, et au niveau ultime on peut faire l'hypothèse qu'ils se sont ainsi réduits car les primates – et les humains plus que d'autres peut-être – se reposent d'avantage sur la vision que sur l'olfaction.

Les explications du niveau ultime reposent sur un équilibre entre les contraintes et les pressions de sélection s'exerçant sur un trait : on pourrait parler de coûts et de bénéfices[16]. On peut interpréter l'idée que le principe de singularité, qui semble limiter drastiquement l'efficacité du raisonnement, est dû en partie aux limites de la mémoire de travail (l'explication avancée par Evans) de deux façons différentes mais compatibles. D'un côté augmenter la mémoire de travail pourrait être extraordinairement coûteux, et de l'autre les bénéfices apportés par le raisonnement pourraient être suffisamment faibles pour qu'ils n'occasionnent que des pressions de sélection très modestes. On pourrait argumenter qu'aucune de ces deux options n'est réellement défendable, mais quelle que soit la ligne qu'Evans choisirait le fait de concevoir le principe de singularité comme une limitation entraîne ce type de questions. Il en va de même pour ce qui est de l'aspect sériel du raisonnement. Si on comprend bien qu'un mécanisme sériel entraîne un goulot d'étranglement qui puisse expliquer le principe de singularité, nulle explication n'est donnée de la cause de cet aspect sériel : on sait que d'autres mécanismes psychologiques fonctionnent en parallèle, pourquoi pas le raisonnement ?

Pour la théorie argumentative, le principe de singularité n'est pas réellement une limitation du raisonnement, il découle plutôt de sa fonction. J'ai comparé plus haut le raisonnement et la recherche visuelle. Dans ce cadre, on peut prendre en compte deux types de sérialité, et chacune reçoit une explication différente : la sérialité de la cible, et celle de la recherche. Par sérialité de la cible j'entends le fait que la recherche vise une seule cible à la fois. Lorsque nous recherchons un objet précis dans notre champ visuel, la fonction même de la mémoire à court terme est de nous permettre de nous focaliser sur cet objet en particulier, au mépris des autres. Etant donné que la recherche a généralement une conséquence comportementale immédiate (se saisir de l'objet recherché par exemple), il ne ferait guère sens de rechercher plusieurs objets en même temps : une fois que l'on s'est fixé un objectif, la fonction de ce type de recherche est précisément de ne pas nous en distraire. Si donc les cibles doivent forcément être considérées de manière sérielle, on peut concevoir que la recherche s'effectue en parallèle : plusieurs objets seraient

[16] D'autres considérations entrent bien sûr en compte. Ainsi, un organisme peut avoir pris une voix évolutive qui n'est qu'un optimum local et qui l'empêche de parvenir à un meilleur optimum. Il ne s'agit pas alors d'un problème de contraintes au sens mentionné ici. Mais la simple dichotomie contraintes / pressions de sélection suffira comme heuristique ici.

considérés simultanément, mais comparés à une même cible. La possibilité même de ce type de recherche parallèle dépend de la capacité du système visuel de former une catégorie unique avec les distracteurs par opposition à la catégorie à laquelle appartient la cible. Plus les distracteurs sont divers, et plus ils sont semblables à la cible, plus une telle opération est difficile (voir Li, 2002).

Comment ceci se traduit-il dans le cas du raisonnement ? Les explications concernant la sérialité de la cible et de la recherche sont différentes. Pour ce qui est de la sérialité de la cible, elle vient du fait que la recherche d'arguments correspond à une demande spécifique, ponctuelle et immédiate. A nouveau, on peut comparer cette recherche à la recherche d'un outil improvisé : lorsque le problème n'est pas de savoir quelle tâche accomplir, mais comment parvenir à accomplir une tâche que l'on s'est fixée, il faut alors pouvoir se concentrer sur cette tâche. Il est normal, dans ces circonstances, de se focaliser entièrement sur une cible. Il est intéressant de comparer ce type d'opération aux autres mécanismes liés à la prise de décision : lorsqu'il s'agit non pas de déterminer comment faire quelque chose, mais ce qu'il faut faire (ce qui est une fonction souvent prêtée au raisonnement), les mécanismes sont alors massivement parallèles (que ce soit au niveau cortical ou infra cortical, voir par exemple Bogacz & Gurney, 2007; Doya, 1999; Redgrave et al., 1999). Donc pour la théorie argumentative le fait que le raisonnement se focalise sur une cible unique est une conséquence directe de la fonction qu'il doit accomplir : trouver un argument afin de convaincre un certain public d'une conclusion donnée.

L'explication de la sérialité des mécanismes de recherche est différente, et tient plus des contraintes évolutives. Pour qu'une recherche parallèle soit possible, il faut que le système en question puisse former facilement des catégories contenant les distracteurs et pas la cible. Pour des raisons mentionnées dans la partie sur le fonctionnement du raisonnement, de tels traits semblent extrêmement improbables dans le cas du raisonnement : l'espace dans lequel il cherche n'est pas organisé autour de catégories qui lui sont immédiatement utiles. Il est donc plus probable qu'il repose principalement sur une recherche sérielle, même si on ne peut pas exclure la possibilité qu'il puisse parfois recourir à certaines catégories et donc mener à bien une recherche en partie parallèle. Et comme les explications en termes de limitations doivent tenir compte d'un équilibre entre les contraintes (ici le fait que les catégories ne sont pas naturellement appropriées) et les bénéfices, nous allons voir que la théorie argumentative explique également pourquoi le raisonnement peut facilement

se contenter d'une solution qui soit assez loin de l'optimalité (et donc que les bénéfices de la recherche d'une solution optimale sont très faibles).

Principe de satisficing

La notion de satisficing a été introduite par Herbert Simon. Il s'agit de l'idée que les processus cognitifs ne parviennent pas à des solutions optimales : il suffit qu'ils satisfassent certains critères de bon fonctionnement, sans rechercher une optimalité qui serait à la fois trop coûteuse et superflue. Evans a cependant en tête une notion de satisficing plus précise, portant sur le fonctionnement même du raisonnement, et il est bon de la citer en entier :

> What I mean by [the notion of satisficing] here is that the (single) mental model that we consider in our hypothetical thinking is evaluated by the analytic system and is accepted unless there is a good reason to reject, modify or replace it. This evaluation may be either casual or more effortful, involving active reasoning or mental simulation. Only if the first hypothesis (possible action etc.) is considered unsatisfactory will another be considered. […] Much more common is a process by which possibilities are considered in turn until one is found that satisfies. (Evans, 2007, p.18)

Il donne ensuite un exemple de prise de décision banale (que va-t-il faire aujourd'hui ?), dans lequel il considère tour à tour différentes possibilités et les teste « en référence avec ses objectifs courants » (Evans, 2007, pp.18-19).

La façon dont il décrit le fonctionnement du principe de satisficing est en très bon accord avec la façon dont j'ai précédemment décrit le fonctionnement du raisonnement comme un mécanisme de filtre considérant diverses possibilités l'une après l'autre (la stratégie 'générer et tester'). On retrouve cependant ici le problème soulevé dans la section précédente : il s'agit simplement pour Evans d'une observation, et il n'explique pas réellement les raisons pour lesquelles le raisonnement devrait fonctionner ainsi. Je vais proposer plusieurs explications, dont à mon avis la première est la seule qu'il ait pu considérer, mais qui n'est guère satisfaisante.

L'explication classique du satisficing est en termes de coûts et, à la limite, de faisabilité, de certains calculs : parvenir à une solution optimale étant très coûteux, voire impossible dans certains cas, les organismes doivent se contenter d'approximations. On voit comment l'application que fait ici Evans du principe de satisficing semble réduire les coûts : plutôt que d'envisager de nombreuses hypothèses, une seule est examinée à la fois. Les choses ne sont cependant pas aussi simples. L'impératif de prendre les coûts en compte ne permet pas de faire de prédictions directes sur la façon dont les mécanismes vont fonctionner. Si on met plusieurs modèles en compétition, il ne suffit pas de dire qu'un modèle est avantageux car il permet de réduire certains coûts par rapport à un maximum donné : les points de comparaison intéressants sont d'autres modèles possibles. Etant donné qu'il n'y a aucune raison de penser que le problème des coûts ait pu être ignoré dans la construction des mécanismes cognitifs autres que le raisonnement, et si la description que fait Evans du satisficing dans le raisonnement découle simplement de considérations de coûts, alors on devrait retrouver des designs semblables pour les autres mécanismes cognitifs. Or ce n'est pas du tout le cas : de nombreux mécanismes cognitifs fonctionnent parfaitement en parallèle, et à moins de faire l'étrange hypothèse qu'ils soient moins bien faits que le raisonnement, il n'y a aucune raison de penser qu'ils sont moins efficaces que le raisonnement dans leur prise en compte des coûts. On peut penser en particulier aux mécanismes de prise de décision qui, généralement, fonctionnent de façon massivement parallèle : si le raisonnement est lui aussi destiné à résoudre des problèmes de prise de décision, comment se fait-il qu'il soit quant à lui tout à fait sériel et qu'il n'examine qu'une hypothèse à la fois ?

On voit que l'explication la plus directe, en termes d'économie de coûts de calcul, est insuffisante : elle sous-détermine complètement le type de solution qui sera adopté, et donc n'explique nullement le fonctionnement précis imputé par Evans au raisonnement. Cette constatation indique cependant une tension intéressante dans les théories du raisonnement lorsque celles-ci sont appliquées à la prise de décision[17] : pourquoi des mécanismes – mécanismes de prise de décision des

[17] Ce qui me semble être le cas général : même les théories comme les modèles mentaux ou la logique mentale qui se concentrent, dans leurs explications et les expériences, sur des problèmes de raisonnement 'purs' mentionnent que les mécanismes étudiés jouent également un rôle central dans la prise de décision (voir Introduction).

systèmes heuristiques et analytiques – qui sont censés accomplir des tâches similaires ont-ils un fonctionnement aussi différent ? Je reviendrai sur ce point dans la section consacrée aux hypothèses évolutionnistes d'Evans.

Une autre explication visant ce principe de satisficing va maintenant être suggérée, explication propre à la théorie argumentative, mais qui ne concerne qu'un de ses aspects. On peut considérer que le principe de satisficing, tel qu'il est décrit par Evans, a deux parties : d'une part l'aspect sériel qui fait qu'une seule possibilité est considérée à la fois, et d'autre part le fait que ces possibilités ne soient pas soumises à un examen très rigoureux. Nous venons de voir que l'aspect sériel réclame des explications, explications que peut apporter la théorie argumentative (voir la partie concernant le principe de singularité). Mais l'aspect sériel, à lui seul, ne suffit pas : on peut très bien concevoir un mécanisme qui examine des possibilités une à une, mais qui, étant particulièrement exigeant, doit en examiner de nombreuses avant d'en trouver une qui lui convienne. Sur ce point, à nouveau, la notion de satisficing n'est utile qu'en tant que notion relative. A moins que l'espace des possibilités ne soit très restreint, aucun mécanisme sériel ne peut envisager toutes les possibilités, ils doivent donc tous, dans une certaines mesure, se prêter au satisficing. Mais Evans va plus loin, et à juste raison : une fois exclue l'interprétation, devenue triviale, selon laquelle le raisonnement ne peut simplement pas prendre en compte toutes les possibilités, et doit donc faire du satisficing à un certain degré, il reste la possibilité que le raisonnement soit un mécanisme qui, *plus que d'autres*, s'engage dans du satisficing. Mais il faut alors expliquer pourquoi le raisonnement se prêterait particulièrement bien au satisficing, ce que ne fait pas Evans ; au contraire, son point de vue général le met dans une position défavorable pour répondre à cette question.

On retrouve ici le problème qui vient d'être mentionné d'une absence de différence claire entre le rôle que jouent les mécanismes des systèmes heuristiques et analytique dans la prise de décision. Si la tâche qu'ils ont à accomplir est globalement similaire – prendre de bonnes décisions – on voit mal pourquoi elle requerrait des degrés de satisficing différents. Pour fournir une explication, on peut chercher dans deux directions : d'une part les coûts de calculs, et d'autre part les coûts et bénéfices des décisions qui sont prises sur la base de ce mécanisme de satisficing. Des coûts de calcul plus élevés doivent entraîner, toutes choses égales par ailleurs, une plus grande tendance au satisficing. A l'inverse, lorsque les erreurs deviennent plus coûteuses, le satisficing devient dangereux. Si on envisage

l'hypothèse, assez répandue, que le raisonnement tend à traiter plus spécifiquement les décisions à prendre dans des situations nouvelles, inédites – par opposition aux mécanismes du système heuristique qui tendraient à être utilisés dans des situations qui ont été habituelles durant l'évolution ou le développement – on peut imaginer que les coûts de calculs soient plus importants dans ces circonstances[18]. Qu'en est-il des coûts et bénéfices des décisions prises dans de telles situations ? D'un côté, on pourrait arguer du fait qu'il y a assez peu de chances que des décisions aux conséquences vraiment coûteuses soient totalement nouvelles, ce qui tendrait à accroître la plausibilité de cette explication pour un plus grand degré de satisficing pour le raisonnement. On peut douter cependant que la plupart des défenseurs de cette hypothèse l'accepterait : pour eux en effet, c'est justement parce que les humains ont rencontré beaucoup de situations nouvelles, et que ces situations étaient très importantes, que le raisonnement a évolué. Les décisions à prendre dans ces situations sont donc au moins aussi importantes, si ce n'est plus, que celles à prendre dans d'autres circonstances. Or, si les décisions sont importantes, elles ne justifient pas un satisficing plus important : au contraire, le raisonnement devrait être plus exigeant dans ses recherches de solutions que ne le sont les autres mécanismes.

La théorie argumentative fournit une réponse à la fois plus précise et mieux défendue à cette question. Pour ce qui est des coûts de calcul, il est établi que l'espace des hypothèses est, dans le cas de la recherche d'arguments, potentiellement immense. Mais il s'agit là d'un argument assez imprécis (car on ne doit pas seulement tenir compte de la taille de l'espace de recherche, mais aussi de la façon dont il est ordonné) et loin d'être décisif. Par contre, la théorie argumentative fait des prédictions plus précises sur les circonstances dans lesquelles le raisonnement devrait être utilisé, et sur les coûts et bénéfices que peuvent entraîner ces situations – prédictions qui concordent avec l'idée qu'il devrait y avoir un très fort degré de satisficing dans le raisonnement. Selon la théorie argumentative, nous raisonnons principalement lorsque nous devons chercher des arguments pour convaincre une autre personne. Les bénéfices d'une telle opération varient considérablement selon l'importance de ce dont nous essayons de convaincre l'autre, et il n'y a pas de raison particulière de penser qu'ils sont soit beaucoup plus soit beaucoup moins importants

[18] L'idée que le raisonnement puisse être consacré aux situations nouvelles sera décortiquée et critiquée plus bas.

que pour d'autres domaines de décisions possible. Par contre, on a de bonnes raisons de penser que les coûts d'un échec tendront à être très faibles, voir nuls. Il y a deux types de risques associés avec l'échec d'une tentative d'argumentation. Le premier est simplement de voir nos efforts ne pas porter leurs fruits : l'autre n'est pas convaincu, et le seul coût a été celui de trouver et de prononcer les arguments – un coût très modeste. Le second, plus rare, est de se voir déprécié par la personne que nous avons essayé de convaincre car l'argument était particulièrement mauvais, ou inapproprié. Ce second problème doit être gardé sous contrôle, mais il n'est pas propre au raisonnement : il s'applique à toutes nos actions, y compris les autres actes communicatifs. Dans tous les cas, une action stupide ou inappropriée peut nous faire mal voir. Dans la mesure où il s'agit d'un problème général (qui s'appliquerait aussi bien au raisonnement quelque soit sa fonction), on peut l'ignorer ici.

Reste donc le premier type de coût, celui associé avec un simple refus de l'argument (ou refus de sa conclusion malgré l'acceptation de l'argument en soi). Pour bien comprendre à quel point ce coût est faible, on peut le comparer aux échecs dans d'autres domaines de la prise de décision : si nos mécanismes de détection de prédateurs ne remplissent pas leur rôle, nous risquons de nous faire dévorer, si nos mécanismes de reconnaissance de nourriture font une erreur, nous pouvons nous empoisonner, etc. Dans l'immense majorité des cas, les environnements physique et biologique sont beaucoup moins cléments que ne l'est l'environnement social. Etant donné que les membres d'un groupe ont tous, dans une plus ou moins grande mesure, besoin les uns des autres, ils ont tous des raisons de ne pas se tuer ou se blesser grièvement. Il y a bien sûr des exceptions, mais, étant donné le nombre d'interactions, elles restent très rares. Ces considérations s'appliquent particulièrement à la communication : si on me dit quelque chose que je n'accepte pas, cela ne constitue pas nécessairement une raison de punir la personne qui me l'a dit, au contraire d'un cas ou cette même personne m'agresserait, ou essaierait de flirter avec mon ou ma partenaire. On a donc de bonnes raisons de penser que dans les contextes argumentatifs, un échec est très peu coûteux : un bon argument en faveur du satisficing.

Un autre élément important en faveur du satisficing dans les contextes argumentatifs est la facilité avec laquelle de nouvelles tentatives peuvent être faites. Dans de nombreuses situations de prise de décision, le premier choix est irrémédiable, ou du moins il entraîne des conséquences importantes. La façon dont

nous nous comportons lorsque nous voyons une personne pour la première fois peut marquer durablement nos relations, et il est parfois dur de remédier à une mauvaise première impression. Si nous prenons la décision de suivre le chemin de gauche, alors que notre proie avait suivi celui de droite, nous rentrerons bredouille. Sans parler de la décision à prendre face à un prédateur… Dans le cas de l'argumentation, un premier échec est presque normal : nous serions parfois surpris de voir l'autre céder au premier argument. Au contraire, si ce premier argument n'était pas assez bon, nous pouvons aisément revenir à la charge, puis tenter encore, jusqu'à ce que nous soyons à court d'arguments. Echouer à trouver le meilleur argument du premier coup n'est presque d'aucune conséquence : l'objectif est de convaincre l'autre et, même si on peut imaginer des gradations dans son degré de conviction, il importe souvent peu de l'avoir convaincu 'largement' ou 'de justesse'. Et si le premier argument échoue, nous pouvons réessayer, en espérant qu'un des arguments suivants sera plus approprié à la situation et au destinataire.

La recherche d'argument est le cadre *idéal* dans lequel la forme de satisficing décrite par Evans – l'examen sériel de possibilités avec application de critères très laxistes – est apte à donner de bons résultats. Cette correspondance me semble être vraiment frappante, d'autant plus qu'Evans est parvenu à cette description du raisonnement sans a priori théoriques la favorisant. Dans la mesure où les propriétés des contextes argumentatifs qui rendent cette forme de satisficing parfaitement approprié leurs sont propres – et j'ai donné plusieurs arguments amenant à penser que les autres situations de prise de décision ne leur ressemblent pas sur ces points – il s'agit là d'un argument assez fort pour la théorie argumentative.

Principe de pertinence

Pour Evans, le principe de pertinence signifie que « les modèles mentaux sont générés pas les processus heuristiques ou pragmatiques qui sont construits pour maximiser la pertinence dans un contexte particulier, étant donné les objectifs du raisonneur. » (Evans, 2007, p.18). Comme il le note lui-même, ce principe est « proche du principe cognitif de pertinence [de Sperber & Wilson, 1995]. ». Prenant ici pour acquis le principe de pertinence, il n'y aura que peu de choses à ajouter. On peut simplement contester un argument que fait Evans dans son ouvrage précédent

de 1996. Il y dit préférer sa version de la pertinence en ce qu'elle serait plus générale car elle « détermine le focus de l'attention des sujets ou le contenu de la pensée. » (Evans & Over, 1996, p.48). On voit mal comment une telle notion de pertinence peut-être plus générale que la notion de pertinence cognitive. Cette dernière s'applique également à tous les processus cognitifs, y compris ceux qui « détermine[nt] le focus de l'attention des sujets ou le contenu de la pensée ». Au contraire, la notion qu'utilise Evans est restreinte aux processus du système heuristique – elle ne conditionnerait le fonctionnement du système analytique que de par les éléments qu'elle lui présente. Evans donne sûrement une telle restriction pour différencier justement le fonctionnement du système analytique, qui est censé être plus éloigné des objectifs courants, plus 'décontextualisé' (pour reprendre le terme de Stanovich).

Mais le principe de pertinence doit continuer de s'appliquer au fonctionnement du système analytique, et ce de deux manières différentes. Au-delà de l'information présentée, la pertinence détermine également l'activation même du système analytique. Cette influence est d'autant plus nécessaire que le système analytique pourrait être activé n'importe quand : nous prenons des décisions et, encore plus, faisons des inférences en permanence, et il ne peut les vérifier toutes – il ne peut même en examiner qu'une infime minorité. Il faut bien que des mécanismes de régulation déterminent les circonstances qui mèneront à son activation. Evans concéderait peut-être cet argument. Mais cela ne suffit pas : la pertinence doit également influer sur le fonctionnement même du raisonnement. Pour ne prendre qu'un exemple, il serait vraiment étrange, et assez dysfonctionnel, d'avoir un niveau de satisficing uniforme à travers toutes les situations : différents contextes appelleront différents niveaux d'exigence. Nous serons par exemple plus regardants vis-à-vis des arguments envisagés en nous adressant à notre jury de thèse qu'à un ami qui vient de payer sa quatrième tournée… Des mécanismes de régulation tels que le principe cognitif de pertinence doivent donc jouer un rôle ici aussi. Le principe cognitif de pertinence, tel qu'énoncé par Sperber et Wilson en 1995 me paraît donc être un outil plus général et qu'il est préférable de conserver tel quel.

Les biais fondamentaux

Selon Evans, les différents principes dont il a expliqué le rôle (principes de singularité, de satisficing et de pertinence) sont à la source de deux « biais fondamentaux » qui seraient responsables des erreurs commises dans des tâches de raisonnement. Le premier de ces biais est le « biais heuristique fondamental », c'est-à-dire le fait que « les gens se focalisent sélectivement sur l'information qui est indiquée comme pertinente », et le second le « biais analytique fondamental », qui correspond au fait que « les gens maintiennent le modèle mental courant avec une évaluation et/ou une considération des alternatives insuffisante » (Evans, 2007, p.22). Le premier découlerait du principe de pertinence, alors que le second serait lui causé par les principes de singularité et de satisficing.

Etant donné qu'on vient de noter l'accord global existant entre la théorie d'Evans et la théorie argumentative sur l'existence (mais pas sur la cause) des trois principes, il serait difficile de refuser ces biais qui se présentent comme en étant des conséquences. A défaut de chercher à réfuter leur existence, il est important d'insister sur le fait que ces biais ne doivent pas forcément être considérés comme des défauts. Le terme de biais à en effet une connotation très négative, et tous les biais ne sont pas également importants. Il est aussi crucial de se demander quels sont les critères qui sont utilisés pour savoir si un mécanisme est biaisé.

Commençons donc par souligner la nécessité, pour pouvoir parler de biais, d'un critère de jugement. Un mécanisme n'est jamais biaisé 'en soi', il ne l'est que par rapport à un certain critère d'exigence, duquel il s'éloigne de façon systématique. Nous avons déjà vu que, dans la nature mais aussi en ingénierie et dans toutes les entreprises ayant une contrepartie matérielle, les considérations de coûts sont aussi importantes que celles de résultats. Or, dans l'usage qui en est habituellement fait, on restreint le terme 'biais' aux résultats. Si on comparaît par exemple deux puces électroniques, une qui ferait des calculs d'une façon particulièrement économique mais qui commettrait certaines erreurs lors du calcul, très rare, des racines $102.354^{\text{èmes}}$, et une autre qui serait parfaite pour la même gamme de calculs, ne commettant jamais aucune erreur, mais qui consommerait une quantité énorme d'énergie, alors c'est la première qui serait jugée comme étant biaisée – bien qu'elle soit ensuite préférée pour toutes (ou presque) applications pratiques. C'est ce type de

considération qui a amené Simon à la notion de satisficing : dans de nombreux cas, on préférera un mécanisme biaisé mais peu coûteux à un mécanisme fonctionnant mieux mais pour un coût disproportionné à l'amélioration apportée. Il est important de se souvenir qu'un mécanisme biaisé peut-être, à toutes fins utiles, supérieur à toutes les variations non biaisées possibles. Dans ce cas, le mécanisme pourrait être 'parfait' pour ce qui est du rapport entre les coûts et les résultats qu'il serait encore biaisé, et alors la connotation très négative de biais semble plutôt malvenue.

Encore faut-il que les erreurs commises par le mécanisme soient systématiques pour qu'elles puissent être qualifiées de biais. Dans quels cas est-ce que des erreurs seront systématiques ? Pour le comprendre, on peut prendre l'exemple d'un jeu de fléchettes fictif. Dans ce jeu, la règle est simplement de viser le centre, seul résultat rapportant des points. Imaginons maintenant un organisme dont la seule fonction soit de viser le centre de ce jeu de fléchettes. Etant donné des ressources limitées, et si la tâche est calibrée pour être assez difficile, l'organisme aura tendance à évoluer vers une distribution aléatoire des erreurs autour du centre (pas vers un biais donc), sauf pour deux types de raisons. Si par exemple l'organisme avait une tendance à viser trop bas (plus d'erreurs en bas qu'en haut), il devrait pouvoir en profiter : en décalant son tir vers le haut, il aurait un meilleur taux de réussite. Nous ne retrouverions alors dans la situation ou les erreurs sont réparties aléatoirement. Qu'est-ce qui pourrait empêcher l'organisme d'effectuer un tel ajustement ? Deux choses. D'une part, cette tendance à viser vers le bas pourrait être due à un processus qu'il serait très coûteux de modifier car il joue un rôle essentiel dans le fonctionnement de l'organisme. D'autre part, cette tendance à viser vers le bas pourrait avoir des bénéfices cachés (demander moins d'énergie car la poussée initiale est moins forte). Dans tous les cas donc, une erreur systématique, un biais, est le résultat d'un compromis qu'on qualifierait, chez un ingénieur, de 'réfléchi'. Ce type d'« erreur » est donc très différent de l'erreur non systématique, résiduelle, et il nous donne des indications possibles sur le fonctionnement de l'organisme.

Des biais peuvent également être créés de toutes pièces lorsqu'on choisit, pour évaluer un mécanisme, des critères de jugement différents de ceux qui ont été retenus par son 'créateur' (la sélection naturelle dans les cas qui nous intéressent ici). Prenons le principe de pertinence. En première approximation on peut estimer qu'il s'agit d'un mécanisme qui régule la façon dont nos pensées et nos actions s'enchaînent. Il est certain qu'il commettra certaines erreurs non systématiques. Il est

également fort probable qu'il soit victime de certains biais. Mais il sera également biaisé par rapport à d'autres critères : il vise à maximiser l'aptitude inclusive des organismes, et pas, par exemple, à les mener à une compréhension profonde du monde. En choisissant des critères externes, qui ne recoupent qu'en partie ce que la sélection naturelle a favorisé, des biais apparaîtront forcément : les chances qu'un mécanisme spécialement adapté pour résoudre une tâche complexe puisse également en résoudre une autre avec exactement la même aisance sont minimes. Nous avons donc ici une deuxième source de biais, qui est cependant beaucoup plus artificielle que la première. Or Evans (et il n'est pas le seul) est relativement équivoque sur les critères qu'il entend utiliser.

C'est un problème car selon l'origine des biais, on ne va pas étudier le phénomène de la même façon. Les biais 'naturels', dus à des compromis dans l'évolution des organismes, peuvent servir de base pour faire des inférences sur la façon dont ces mécanismes fonctionnent, sur les compromis qui ont dû être faits durant leur évolution. La difficulté dans ce cas est d'établir le critère qui servira à mesurer les biais. Dans une perspective évolutionniste, il s'agit de la fonction attribuée au mécanisme en question, mais celle-ci sera souvent difficile à déterminer. Cependant ce même problème se pose pour les biais 'artificiels', ceux qui résultent de l'utilisation de critères différents de ceux retenus par la sélection naturelle. Afin que les données sur ces biais artificiels puissent être exploitées pour faire des inférences sur le fonctionnement des mécanismes psychologiques sous-jacents, une première étape est de pouvoir les différencier des biais naturels. Lorsqu'on observe une déviation systématique par rapport à une norme artificielle, on ne peut pas immédiatement savoir quelle part est due à cette artificialité même et quelle part appartient aux biais naturels des mécanismes étudiés. Il convient donc de commencer par soustraire les biais naturels, afin de parvenir aux biais artificiels. Mais une telle opération implique à nouveau de connaître la fonction des mécanismes étudiés.

Notons qu'on pourrait faire un parallèle intéressant entre ces deux types de biais et les deux types de rationalités qu'Evans et Over proposent. A la rationalité1, celle qui répond aux besoins de l'organisme, correspondraient alors les biais naturels, et à la rationalité2, celle qui nous fait respecter des normes telles que celles de la logique, correspondraient les biais artificiels. Le problème est que cette convergence ne se prolonge pas au niveau des mécanismes censés sous-tendre ces deux types de rationalité. Ainsi, le système heuristique est censé être 'responsable' de la

rationalité1, mais il arrive souvent qu'on le mesure à l'aune de normes tout à fait artificielles – auquel cas on ne peut pas savoir quelle est la part de naturel et d'artificiel dans les biais résultants. De même, le système analytique est censé sous-tendre la rationalité2. Le problème, c'est que si la fonction du système analytique est précisément de nous permettre de suivre ces normes (et c'est l'opinion d'Evans, semble-t-il), les biais résultants seraient alors au contraire des biais 'naturels' – puisque dans ce cas la norme ne serait plus artificielle.

Un moyen assez grossier, mais qui me semble être assez fiable, est d'utiliser l'ampleur du biais pour découvrir s'il s'agit d'un biais naturel ou artificiel. Plus le biais est important, plus il a de chances d'être artificiel. En effet, plus un biais est important, plus il faut qu'il existe, par ailleurs, de bonnes raisons de le conserver pour qu'il soit effectivement conservé. Or ces conditions ont de moins en moins de chances d'être réunies lorsqu'elles deviennent plus importantes : un biais qui ferait systématiquement confondre proies et prédateurs requerrait par exemple de très bonnes raisons pour expliquer sa présence ! On peut donc utiliser la force d'un biais pour former des hypothèses sur ce qui est (ou plutôt sur ce qui n'est pas) une fonction plausible pour le mécanisme étudié. Si on fait l'hypothèse qu'un mécanisme dessert une certaine fonction, mais qu'on se rend compte que les réponses qu'il donne son extrêmement biaisées par rapport aux critères qui découlent de sa fonction supposée, c'est une bonne raison de penser qu'il s'agit de biais artificiels, c'est-à-dire que les critères choisis sont artificiels, et qu'on peut donc rejeter cette fonction. Il est alors possible de chercher une fonction pour laquelle il s'agira en effet de biais artificiels et non de biais naturels (nous en verrons un exemple dans la discussion du biais de croyance). Cette façon de procéder permet de faire des hypothèses de plus en plus plausibles sur la fonction des mécanismes étudiés. Notons pour finir que toute cette discussion nécessite, pour être valide, un point de vue résolument adaptationniste. Hors de ce cadre les seules normes possibles sont les normes artificielles que l'on peut vouloir imposer, la distinction entre biais naturel et artificiel disparaît, et il devient dès lors plus difficile de faire des inférences utiles sur le fonctionnement des mécanismes psychologiques sous-jacents.

L'interaction entre les deux systèmes

La quatrième direction dans laquelle Evans a précisé sa théorie concerne les interactions entre les deux systèmes. Voici la façon dont, selon lui, les choses se passent. Une décision ou un jugement est rendu par le système heuristique. Le système analytique a alors au moins « une implication minimale, même si ce n'est que pour approuver (peut-être sans réflexion) la réponse par défaut suggérée par les modèles mentaux dérivés de façon heuristique. » (Evans, 2007, p.20). Trois facteurs déterminent le degré d'implication du système analytique, et donc les chances qu'il modifie la réponse par défaut du système heuristique : les instructions, l'intelligence générale et le temps disponible. Plus les instructions mettent l'accent sur la nécessité de réponses logiques, plus les participants sont 'intelligents', et plus ils ont de temps pour répondre, plus le système analytique pourra faire son œuvre et, le cas échéant, modifier la réponse donnée par le système heuristique.

C'est probablement à ce niveau que la théorie d'Evans et la théorie argumentative font les prédictions les plus différentes. Les prédictions de la théorie argumentative sur les contextes promouvant l'utilisation du raisonnement seront présentées plus loin (voir l'introduction de la partie empirique). Ici je me concentrerai sur les limites des propositions d'Evans. Avant d'en venir aux conditions censées favoriser l'activation du système analytique, on peut s'interroger sur l'idée qu'il serait en fonction en permanence, passant en revue tous les jugements et décisions du système heuristique. Cette idée (pourtant partagée par Kahneman & Frederick, 2002) paraît très peu plausible, à moins de spécifier de façon ad hoc les jugements et décisions concernés. Il est assez couramment admis, y compris par Evans ou d'autres partisans des théories à processus duel, que les mécanismes du système heuristique fonctionnent de façon massivement parallèle : de nombreux traitements peuvent être effectués au même instant. Imaginons par exemple que vous soyez en train de parler avec quelqu'un. En même temps que des processus consacrés à la production d'énoncés sont activés, vous prêtez attention aux réactions de l'interlocuteur, à son langage corporel, à ses expressions, pour vous former une idée de la façon dont vos propos sont reçus. Et il ne s'agit là que de la pointe de l'iceberg : pendant ce temps, votre cerveau mène à bien de nombreuses autres opérations : votre hippocampe stocke certains faits en mémoire, votre amygdale

estime la valence émotionnelle du contexte, etc. Il serait vraiment très étonnant – en fait, computationnellement impossible – que le système analytique puisse examiner tous ces éléments simultanément. Il y a même une contradiction directe entre l'aspect massivement parallèle des traitements du système heuristique, le principe de singularité régulant la capacité du système analytique et l'idée qu'il puisse examiner tous les jugements et les décisions qui sont pris. La seule solution serait de restreindre artificiellement la notion de 'jugement' et de 'décision' à ces jugements et décisions qui sont supervisés par le système analytique, mais alors dire qu'il les supervise tous devient tautologique.

Passons maintenant en revue les critères censés déterminer le degré d'activation, ou d'exigence, du système analytique. Le premier est le jeu d'instructions fourni au participant : plus il accentue la logique, plus la performance des participants est bonne. Etant donné qu'une bonne performance est mesurée à l'aune de critères logiques, on ne sera guère étonné par cette observation. Mais son aspect trivial n'est pas son plus gros défaut : elle manque également de validité écologique. Il s'agit là d'un triste reflet, d'un exemple presque caricatural, de la tendance qu'ont les psychologues de construire des théories qui ne s'appliquent qu'aux tâches qu'ils ont fabriquées eux-mêmes. Lorsque nous prenons des décisions ou formons des jugements dans la vie de tous les jours, ce n'est pas pour répondre à un ensemble d'instructions. Il est vraiment dommage qu'Evans ne tente même pas de tirer des leçons plus générales de cette observation. On pourrait imaginer par exemple qu'elle reflète la volonté des participants de suivre les demandes de l'expérimentateur, ou qu'elle est le résultat de l'artificialité des normes logiques – car il faut réellement insister pour qu'elles soient prises en compte. Quoiqu'il en soit, en l'état, ce critère n'est d'aucune utilité pour savoir quand les gens raisonnent dans la vie de tous les jours.

Le second critère, l'intelligence générale, n'est guère plus utile. S'il peut, peut-être, expliquer certaines des différences entre individus dans la tendance à raisonner (à utiliser leurs systèmes analytique), il ne peut pas expliquer pourquoi un même individu raisonne à un moment et pas à un autre, à moins de postuler des différences intra-individuelles dans l'intelligence générale, qui existent sûrement en effet mais réclament explication à leur tour. De plus, dans une perspective évolutionniste, les mécanismes psychologiques qui sont considérés comme des adaptations – ce qui est certainement le cas du raisonnement pour la théorie

argumentative – doivent être universels et ne peuvent que peu varier entre individus : c'est là l'effet même de la sélection naturelle[19]. Invoquer des différences interindividuelles pour expliquer l'activation différentielle d'un mécanisme est alors très limité. On imagine mal, par exemple, expliquer l'activation du mécanisme de détection des visages non par la présence d'un visage dans le champ de l'attention mais par le fait que certaines personnes sont meilleures que d'autres à détecter les visages. Ou plutôt, une telle explication ne serait invoquée que dans des cas de déficits : on pourrait par exemple expliquer une absence d'activation par une lésion dans une zone de traitement visuel antérieure. Mais dans le cas du raisonnement ce ne sont pas, dans la littérature qui nous occupe ici, des cas de lésions ou des déficiences profondes qui sont étudiées. Quoiqu'on pense des étudiants dans les premières années d'université, on ne peut pas leur prêter de retard mental profond.

Enfin, le troisième critère – le temps disponible – ne fournit que des indications fort vagues. Dans le cadre strictement expérimental, il permet en effet de faire des prédictions testables sur les mécanismes qui seront utilisés en fonction du temps de réponse imparti aux participants (voir Evans & Curtis-Holmes, 2005, par ex.). Mais sa validité écologique est limitée : comment peut-on l'appliquer en dehors de ces contextes ? La seule chose qu'il puisse faire est de fournir une borne inférieure à l'activation du système analytique : celui-ci ne pourrait pas être utilisé lorsque le temps pour prendre une décision est vraiment trop court. Mais même dans ce cas, ce critère ne semble pas bien fonctionner : lors d'un débat par exemple, deux interlocuteurs peuvent échanger des arguments très rapidement, or ils ont dû user du raisonnement pour trouver de tels arguments. On pourrait alors arguer du fait que même dans ce cas il existe une certaine marge : on peut toujours prendre quelques secondes avant de répondre à une objection. Mais si la limite posée par le critère de temps disponible ne s'applique qu'aux décisions qu'il faut prendre en un temps très court – inférieur à la seconde par exemple – alors il est vraiment faible : nous avons presque toujours plus de temps disponible pour prendre nos décisions. Si donc nous avons suffisamment de temps, dans une grande majorité des cas de prise de décision (et de jugement), pour raisonner, alors ce critère ne nous est guère utile.

[19] A quelques exceptions près (sélection sexuelle et sélection fréquence dépendante) qu'il est difficile d'envisager pour le raisonnement, et qui feraient des prédictions totalement différentes sur la répartition des traits phénotypiques.

Si on accepte que les matériaux sur lesquels le système analytique travaille sont extrêmement nombreux, que se soit à un moment donné ou de façon successive, le besoin de critères beaucoup plus contraignants que ceux proposés pas Evans apparaît clairement. Les critères proposés par la théorie argumentative seront évoqués plus loin. Ici, je me contenterai de remarquer une absence surprenante chez Evans (mais qui ne lui est pas propre) : le critère de pertinence.

Lorsqu'Evans mentionne la pertinence, c'est pour expliquer quel modèle est considéré par le système analytique. Il ne mentionne pas qu'elle puisse jouer un rôle dans l'activation même de ce système. Or c'est justement un des résultats centraux des théories à processus duel dans le domaine de la persuasion et du changement d'attitude. De nombreuses études (à commencer par Petty & Cacioppo, 1979, voir Petty & Wegener, 1998, pour revue) ont en effet montré que lorsque les participants examinent un argument dont la conclusion est pertinente pour eux, ils auront plus tendance à examiner le fond de l'argument que des critères superficiels (comme le charme de la personne l'énonçant). Or Evans conviendrait sûrement que faire la différence entre les bons et les mauvais arguments nécessite l'utilisation du système analytique. On peut alors s'interroger sur l'absence de ce critère simple de pertinence parmi ceux envisagés par Evans. Ici encore, on peut redouter que ça ne soit l'effet d'une focalisation exclusive sur les problèmes artificiels de la psychologie du raisonnement, problèmes dont les conclusions ne sont *jamais* pertinentes pour les participants – ce facteur ne variant pas, on ne peut guère en observer les effets. Plus généralement, on ne peut que regretter que l'intégration entre les différentes théories à processus duel ne reste souvent qu'une déclaration de principe, et que les résultats centraux d'un domaine soient totalement ignorés par les autres. Cette absence d'intégration est d'autant plus dommageable que les domaines du changement d'attitude et de la psychologie du raisonnement ont de nombreux points communs, qu'il s'agisse des problèmes utilisés (évaluation d'arguments) ou des théories dominantes (à processus duel).

Hypothèses évolutionnistes

Evans n'accorde guère d'importance aux théories évolutionnistes : « Bien que ces spéculations [évolutionnistes] soient intéressantes, elles semblent n'avoir que peu de pertinence pour les chercheurs étudiant la pensée et le raisonnement qui tentent de rendre compte des résultats de leurs expériences » (Evans, 2006). On ne pourra donc guère lui tenir rigueur de ne pas avoir réellement développé d'hypothèses précises sur les raisons évolutives de l'apparition des mécanismes du système analytique. La perspective dans laquelle la théorie présentée ici se situe est cependant résolument évolutionniste, et il est donc utile d'examiner les quelques éléments qui sont fournis par Evans.

Le premier est un point sur lequel les deux théories sont en accord : il partage l'opinion défendue tout d'abord dans le domaine de l'apprentissage implicite que le système heuristique est plus ancien évolutionnairement que le système analytique, qui serait lui apparu très récemment. Cette hypothèse ne nous amène cependant pas bien loin. Dans l'ouvrage écrit avec Over, ils donnent un peu plus de précisions sur ce qui pourrait être la fonction des mécanismes analytiques : « L'avantage d'un système à processus duel est que la réflexion consciente [le système analytique] fournit la flexibilité et la prévoyance que le système tacite [le système heuristique] ne peut pas, de par sa nature même, fournir. » Ou encore, plus loin : « la conscience [le système analytique] nous donne la possibilité de traiter les nouvelles situations et d'anticiper le futur » (Evans & Over, 1996, p.154). On retrouve une idée similaire chez d'autres partisans des théories à processus duel. Steven Sloman cite même un passage de William James disant exactement la même chose : « Le vrai raisonnement est "productif", selon James, car il peut traiter de nouvelles données : "Le raisonnement aide à nous tirer de situations sans précédents" » (Sloman, 1996, p.4, citant James, 1950, p.330)[20]. Cette idée n'est donc pas nouvelle, et la critique qui sera adressée à cette position ne se limite pas à la version défendue par Evans mais s'étend aux théories apparentées.

Le problème général de cette vision du raisonnement en tant que mécanisme permettant de traiter de la nouveauté est qu'il résulte d'une confusion avec la

[20] Une idée nullement propre au monde anglo-saxon : selon Claparède, l'intelligence est un « processus destiné à résoudre par la pensée un problème nouveau » (Claparède, 1923)

fonction de l'apprentissage en général. Deux chercheurs spécialistes de l'apprentissage en donnent la définition suivante : « L'apprentissage, sous sa forme la plus basique, peut être vu comme un processus par lequel nous devenons capables d'utiliser les événements courants et passés pour prédire ce que réserve le futur. » (Niv & Schoenbaum, 2008). Si nous étions confrontés tout le temps à la même situation, il n'y aurait nul besoin d'apprentissage. L'apprentissage nous sert précisément à tirer profit de nos erreurs (et de nos succès) pour savoir ce qu'il faut faire lorsqu'une situation, qui était nouvelle la première fois, se présente à nouveau, ou qu'une situation similaire se présente. On pourrait alors rétorquer que dans ce cas l'apprentissage ne sert qu'une fois que la situation n'est plus nouvelle, lorsqu'on la rencontre à nouveau. Soit. Mais le problème est alors que si une situation est entièrement nouvelle, qu'aucune réaction appropriée à une situation s'en approchant n'est connue[21], alors il est *impossible* d'avoir une réaction qui tende à être meilleure que le hasard.

Le problème est encore plus frappant si on considère des modèles d'apprentissage qui dépassent les modèles élémentaires de conditionnement. Le modèle le plus utilisé est celui de Rescorla et Wagner (Rescorla & Wagner, 1972). Bien qu'il tende maintenant à être remplacé par des modèles plus sophistiqués, il permet tout de même de rendre compte d'un grand nombre de données, et illustrera parfaitement mon argument (qui reste valide, qui l'est même plus encore, avec les modèles plus récents). Dans ce modèle, les organismes apprennent en prédisant le futur, en notant la différence entre ce qui se passe et la prédiction qu'ils avaient faite, et en utilisant cette différence pour mettre à jour leur base de connaissance (qui à son tour leur permet de faire de nouvelles prédictions). On retrouve ici tous les éléments requis par Evans et Over : l'utilisation des connaissances actuelles afin de prédire le futur dans des situations qui sont nécessairement en partie nouvelles (sinon le processus dans son ensemble n'aurait aucune utilité car son taux de succès serait de 100%). Il semble que rien ne différencie la tâche qu'accompli un tel mécanisme – bien connu, et dont les bases cérébrales sont étudiées chez les rats aussi bien que chez les humains – et celle qui est prêtée au raisonnement par ces auteurs.

La fréquence à laquelle on retrouve cette intuition, selon laquelle le raisonnement serait principalement utile face à des situations nouvelles, fait qu'elle

[21] En incluant les mécanismes ayant des bases innées.

réclame tout de même une explication. Des éléments seront fournis plus loin expliquant pourquoi le raisonnement tend à être activé plus facilement dans un type bien précis de situations qu'on peut qualifier de 'nouvelles'. Des hypothèses seront également avancées pour rendre compte du fait qu'il peut, en effet, amener à de meilleures décisions dans ces circonstances (même si c'est loin d'être toujours le cas). Ces deux explications découleront de la façon dont le raisonnement fonctionne selon la théorie argumentative (et d'autres considérations), mais je ferai alors face à un dilemme : si le raisonnement, admettant même que sa fonction première soit argumentative, permet en effet de prendre de meilleures décisions dans certains cas (ce qui est indéniable), alors pourquoi n'a-t-il pas pu évoluer, ou du moins être coopté, pour cette tâche ?

Il n'est pas possible, j'en ai peur, de faire une réponse catégorique à cette question, et la solution proposée dépend de considérations de plausibilité, et de coûts et de bénéfices relatifs. On peut penser, de prime abord, que la simple démonstration qu'un mécanisme psychologique peut apporter certains bénéfices (prendre de meilleures décisions grâce au raisonnement dans le cas qui nous occupe) suffit à prouver qu'il s'agit d'une des raisons pour lesquelles il a été, ou est, sélectionné. En effet, on pourrait penser que les chances qu'un mécanisme donné, ayant une fonction spécifique, fonctionne bien pour un autre domaine sans qu'il n'ait été sélectionné pour sont relativement maigres. Ce n'est cependant pas nécessairement le cas. Il arrive parfois qu'un mécanisme assez général soit plus efficace qu'un mécanisme plus spécifique. C'est par exemple le cas de la vision en couleur. Elle n'a pas évolué pour qu'on puisse distinguer les nuances fines d'un coucher de soleil sur une mer turquoise, mais pour reconnaître les fruits mûrs. Il était cependant plus simple d'avoir un mécanisme qui voie tout en couleur qu'un mécanisme qui ne voie que les fruits en couleur. Il aurait fallu dans ce cas rajouter une 'bride' coûteuse et inutile. Je suggère que le raisonnement est un cas similaire : le fait qu'il puisse parfois nous aider à prendre de meilleures décisions peut entièrement s'expliquer par une heureuse coïncidence entre ce que sa fonction (trouver des arguments) lui permet de faire et la structure de certains problèmes pour lesquels il n'était pas fait.

Mais cela ne fait qu'ouvrir la possibilité que le raisonnement ait pu ne pas être sélectionné pour cette fonction de prise de décision. Il faut d'autres arguments pour montrer qu'il ne l'a, en effet, pas été. J'utiliserai pour cela une question portant sur un contrefactuel : existe-t-il un mécanisme plus économique, ou plus efficace,

que le raisonnement qui pourrait parvenir aux mêmes résultats (dans le domaine qui nous occupe ici, à savoir la prise de décision) ? Si on peut envisager l'existence d'un tel mécanisme, mais qu'on n'observe cependant pas sa présence, on peut conclure que les bénéfices qu'il apporterait ne valent pas les coûts qu'il occasionnerait, et, a fortiori, qu'aucun autre mécanisme aux coûts supérieurs n'a pu être sélectionné pour la même tâche. Il faudrait donc pouvoir montrer deux choses : premièrement qu'un mécanisme alternatif au raisonnement pourrait avoir, globalement, les mêmes effets (pour ce qui de prendre de meilleures décisions) et deuxièmement qu'il engagerait, dans l'accomplissement de sa tâche, moins de dépenses que le raisonnement.

On peut parvenir à un tel résultat en partant uniquement d'éléments couramment admis dans les théories à processus duel. Les mécanismes intuitifs y sont généralement vus comme étant plus efficaces que le raisonnement lorsqu'ils font face à des situations qu'ils 'connaissent' (soit car ils ont appris à les traiter durant le développement, soit qu'il existe des bases innées acquises durant l'évolution). Le raisonnement serait efficace au contraire dans les situations qui s'éloignent de celles-ci. Par ailleurs, le fait que le raisonnement est plus coûteux que les mécanismes intuitifs est largement reconnu (voir Gailliot et al., 2007; Masicampo & Baumeister, 2008 pour des éléments empiriques). Mais alors, pourquoi est-ce que de nouveaux mécanismes intuitifs, plus efficaces que le raisonnement si ce n'est que parce que moins coûteux, n'ont pas évolué pour résoudre ces problèmes, faire face à ces situations ?

S'il n'est pas décisif, cet argument a le mérite de renverser la charge de la preuve. Pour montrer que le raisonnement a en effet été sélectionné car il nous permet, dans certains cas, de prendre de meilleures décisions, il faut expliquer pourquoi les problèmes posés pas ces situations ne pourraient pas être résolus par de nouveaux mécanismes intuitifs. Dire que les situations dans lesquelles le raisonnement est efficace sont spéciales car 'nouvelles' n'aide pas : avant que des mécanismes dédiés n'y soient consacrés, toutes les situations sont 'nouvelles'.

Pour ceux que ces arguments ne convaincraient pas, on peut se contenter de souligner l'extrême vague dans lequel nous laisse le fait de penser que la fonction du raisonnement est de 'traiter la nouveauté'. Comment faire des prédictions sur le fonctionnement du raisonnement sur une telle base ? A leur crédit, les partisans de

cette idée ne s'y risquent guère, et s'ils ont fait un peu d'ingénierie inverse[22] pour attribuer cette fonction au raisonnement, ils ne se livrent pas à l'exercice suivant de 'penser adaptatif', ce qui est fort dommage.

Conclusion : retour sur la notion de rationalité et sur le rôle des hypothèses évolutionnistes

Pour conclure cette section portant principalement sur la théorie d'Evans, on peut revenir rapidement sur la notion de rationalité, et sur l'importance des hypothèses évolutionnistes. Nous avons vu qu'Evans et Over ont proposé l'utilisation de deux notions de rationalité différentes, une étant censée s'appliquer au système heuristique, et l'autre au système analytique. Voici la définition qu'en donnaient Evans et Over. La Rationalité1correspond à « Penser, parler, raisonner, prendre une décision ou agir d'une façon qui est généralement fiable et efficace pour achever ses propres objectifs » ; et la Rationalité2 à « Penser, parler, raisonner, prendre une décision ou agir lorsqu'on a une raison pour ce que l'on fait sanctionnée par une théorie normative. » (Evans & Over, 1996, p.8). Il semble qu'Evans et Over veulent donner des critères permettant d'évaluer les *effets* de mécanismes des deux systèmes. Pour reprendre la conclusion de la discussion portant sur les biais, il semble que le premier type de rationalité corresponde à des critères naturels quand le second correspondrait à des critères artificiels. Je voudrais maintenant questionner l'utilité même de ces seconds critères, au moins pour ce qui est de la compréhension des mécanismes psychologiques.

Pour comprendre le rôle que peuvent jouer ces critères artificiels, on peut prendre le cas, très similaire à la psychologie dans son objet d'étude, de l'éthologie. Si un courant d'éthologie a toujours porté grande attention à l'écologie et à l'évolution des animaux étudiés, ce ne fut guère le cas de l'éthologie cognitive à ses débuts. Les tâches utilisées étaient alors, en grande majorité, des tâches que nous qualifierons d'abstraites pour des humains et qui ne devaient l'être que plus pour d'autres animaux. Cependant, on observe depuis quelques années une recrudescence de travaux prenant justement en compte les tâches que les sujets doivent remplir dans

[22] C'est-à-dire partir des caractéristiques observées d'un système pour essayer d'en inférer la fonction.

leur environnement naturel, celles pour lesquelles ils sont faits (voir par exemple Bovet, in prep). L'adoption de ces nouvelles tâches, et des nouveaux critères de réussite qui les accompagnent, a donné lieu à des travaux splendides qui, tout en mettant en lumière les capacités extraordinaires de certains animaux, offrent une bien meilleure compréhension de leurs mécanismes psychologiques (voir par exemple Hare, Call, & Tomasello, 2001; Hare, Call, & Tomasello, 2006; Hare & Tomasello, 2004, pour le cas de la cognition sociale). Il n'y a pas de raison majeure pour laquelle la psychologie ne pourrait pas, ne devrait pas même, profiter des leçons de l'éthologie. Bien entendu, les psychologues travaillant dans une perspective évolutionniste et/ou écologique tentent de promouvoir cette idée depuis des années (voir par exemple Gigerenzer, 2007; Tooby & Cosmides, 1992). Il ne faut pas pour autant rejeter entièrement les critères artificiels. Ils peuvent être fort utiles lorsqu'il s'agit d'éducation par exemple : étant donné que ce qui nous est appris à l'école tend à dépasser nos intuitions, il faut bien utiliser des critères nouveaux pour pouvoir évaluer les apprenants. Mais il faut être méfiant lorsqu'on tente de s'en servir pour mieux comprendre le fonctionnement des mécanismes psychologiques dont on pense qu'ils sont le produit de l'évolution.

On pourrait dire, au crédit d'Evans et Over, que dans leur cas les critères posés par la rationalité2 sont, en fait, des critères naturels : respecter des théories normatives serait, pour eux, une des fonctions du système 2. Mais ils sont alors confrontés à un problème : « Nous *pouvons* prendre des décisions conscientes basées sur l'analyse d'un problème nouveau, la projection de modèles mentaux de futurs mondes possibles, et un calcul des risques, des coûts et des bénéfices. Entendu, nous ne sommes pas très bons en prise de décision consciente, de la même façon que nous ne sommes pas très bons en raisonnement déductif, à cause de contraintes cognitives sévères. » (Evans & Over, 1996, p.154). Dans un cadre évolutionniste (dans lequel ils se placent eux-mêmes à cette occasion), un tel aveu est extrêmement étrange. On imagine mal un biologiste dire que « la fonction du système auditif des chiens est de comprendre ce que dit leur maître, mais ils n'y excellent pas, à cause de contraintes cognitives ». Au contraire, lorsqu'un mécanisme donné n'est pas très bon (et dans le cas du raisonnement, certains diraient vraiment mauvais) pour accomplir la fonction qu'on lui a tout d'abord attribuée, l'explication la plus naturelle est simplement que cette fonction n'est pas la bonne. Il y a deux raisons pour préférer une explication en termes de contraintes. Soit on sait qu'il existe des contraintes très puissantes qui ne

peuvent pas être modifiées – car elles résultent de lois physiques par exemple. Soit on a une très forte présomption que la fonction attribuée initialement est en fait la bonne. Or, même si, à première vue, ces deux éléments semblent être réunis dans le cas du raisonnement, on peut montrer que ce n'est pas en fait le cas.

Pour ce qui est des contraintes (telles que la limite de la mémoire de travail), nous ne faisons qu'observer leur présence. De telles observations ne sont pas suffisante pour qu'on puisse les considérer comme des contraintes qui expliquent les limites de l'aptitude : il faut également donner des raisons de penser que les repousser entraînerait des coûts très élevés, et personne, je pense, n'a fait de telle démonstration. On pourrait par exemple imaginer des biologistes expliquant la fréquence de la répartition en trois segments corporels chez les arthropodes par des contraintes, jusqu'à ce qu'ils découvrent les mille-pattes qui les forcent à chercher des explications fonctionnelles. Lorsqu'un trait n'a pas d'explication fonctionnelle (pour une théorie donnée), il est facile de le traiter comme une simple contrainte. Par ailleurs, ce qui peut être vu comme une contrainte pour une théorie peut être le résultat du fonctionnement normal pour une autre. Dans ce cas, il convient de favoriser la seconde théorie, car elle prédit l'existence de ce qui n'est alors plus une contrainte, au contraire de la première qui ne fait que la constater[23]. C'est le cas du principe de singularité : vu comme une contrainte pour la théorie d'Evans, il découle du fonctionnement normal, efficace, du raisonnement pour la théorie argumentative. Puisque la théorie argumentative en fournit une explication en termes de fonctions plutôt que de l'expliquer en termes de contraintes qu'on ne fait que constater, elle a ici un avantage.

Qu'en est-il de la fonction du raisonnement ? A-t-on de fortes raisons, a priori, de penser que le raisonnement a une visée principalement individuelle telle que de nous aider à prendre de meilleures décisions dans des situations nouvelles, ou d'améliorer notre condition épistémique ? Il est dur de nier que cette idée est très intuitive, tant elle est partagée – et semble l'avoir toujours été. Mais il ne s'agit pas là

[23] Et cela a également été observé par des chercheurs ne se plaçant pas explicitement dans une perspective évolutionniste : « There is a sense in which rational explanations are more satisfying than mechanistic explanations. A mechanistic explanation treats the configuration of mechanisms as arbitrary. The justification for the mechanisms is that they fit the facts at hand. There is no explanation for why they have the form they do rather than an alternative form. In contrast, a rational explanation tells why the mind does what it does » (J. R. Anderson, 1991, p.410)

d'une bonne raison, ou d'une raison suffisante : nos intuitions sur la fonction de divers mécanismes ne sont que ça, des intuitions – et l'on sait à quel point elles peuvent être trompeuses en science. Dans la lignée des premiers psychologues évolutionnistes, il me semble préférable, lorsque c'est possible, de remplacer ces intuitions par des considérations basées sur la théorie de l'évolution (Tooby & Cosmides, 1992). Pour ce qui est des capacités cognitives humaines, le courant le plus fort depuis maintenant plus de vingt ans met plutôt l'accent sur l'importance de l'environnement social (Byrne & Whiten, 1988; Whiten & Byrne, 1997). Il serait, plus que d'autres, à l'origine des pressions de sélection ayant façonné notre esprit. Dans ce cadre, c'est au contraire une hypothèse sociale – comme la théorie argumentative – qui devrait tendre à être favorisée. Finalement, il est possible que les intuitions qui ont mené tant d'éminents penseurs à voir dans le raisonnement un outil de la cognition individuelle soient principalement basées sur des artefacts, des éléments créés par la culture, tels que la philosophie, les lois, ou la science. Etant donné que le raisonnement n'a pas pu évoluer à de telles fins, qui ne sont apparues que beaucoup trop récemment, il s'agit là d'une raison de plus d'ignorer nos intuitions concernant la fonction du raisonnement.

Il est donc, d'une part, difficile de faire un argument convaincant pour la force des contraintes étant supposées peser sur le raisonnement et, d'autre part, facile de mettre en doute la fonction qui lui est attribuée par Evans et Over. Il semble donc que la conclusion logique de l'observation des profondes carences du raisonnement lorsqu'il s'agit de mener à bien sa fonction supposée doivent aboutir à la remise en cause de cette fonction et la recherche d'une autre possibilité. Mais avant de considérer que cet argument favorise spécifiquement la théorie argumentative, il faut montrer que les théories alternatives rencontrent, elles aussi, des problèmes.

3.2 La théorie de Sloman

La distinction entre systèmes associationiste et basé sur des règles

Alors qu'Evans et Over mettaient au point la théorie qui sera exposée dans leur ouvrage de 1996, Steven Sloman préparait sa propre version, qui sera publiée la même année (Sloman, 1996). Il s'appuie sur la distinction entre mécanismes associationnistes, qui sont généralement modélisés par le biais de réseaux de neurones, et mécanismes basés sur des règles, qui sont principalement, mais non exclusivement, le fief de l'intelligence artificielle classique :

> Le débat a fait rage (encore) en psychologie cognitive depuis près d'une décade aujourd'hui. Il oppose ceux qui préfèrent que les modèles des phénomènes mentaux soient construits de réseaux de mécanismes associationistes qui transmettent des activation aux alentours en parallèle et sous forme distribuée (de la façon dont les cerveaux fonctionnent probablement) contre ceux qui préfèrent que les modèles soient construits à partir de langages formels dans lesquels les symboles sont assemblés en propositions qui sont traitées séquentiellement (ainsi que fonctionnent les ordinateurs). (Ibid., p.3)

Sloman généralise cette distinction de la façon suivante. D'un côté on trouverait un système associationiste qui « encode et traite des régularités statistiques de son environnement, fréquences et corrélations entre les différentes caractéristiques du monde » (p.4). On retrouve là la vieille psychologie behavioriste, sa crédibilité redorée par d'élégants réseaux de neurones (en oubliant, notons le au passage, qu'elle a été discrédité pas tant parce qu'on ne pouvait la modéliser que parce qu'elle ne pouvait rendre compte de comportements même assez élémentaires). L'autre système serait composé de règles, des « abstractions qui s'appliquent à toutes les propositions qui ont une certaine structure symbolique bien spécifiée » (p.5). Il cite comme exemple de règle la règle de conjonction qui, en probabilité, dit que $P(A) \geq P(A \& B)$ (avec P(A) probabilité de l'événement A). On voit que cette règle s'applique à toutes les paires d'événements possibles.

Bien que la plupart des autres théories à processus duel reprennent cette distinction à leur compte, il me semble que cela pose problème. Je ne pense pas, en effet, qu'elle puisse proprement départager deux systèmes, deux ensembles de mécanismes. Notre objectif ne sera cependant pas de montrer que la distinction proposée par Sloman, à la suite de plusieurs chercheurs dans le domaine de l'intelligence artificielle, ne reflète pas une certaine réalité, mais qu'elle est d'un type très différent de celle qui est habituellement entendue lorsqu'on parle de système 1 et de système 2. Or, les résultats que Sloman veut expliquer à l'aide de sa théorie sont, globalement, les mêmes que ceux que les autres théories prétendent expliquer. Il y a donc ici un problème : les deux types de théories, si elles sont en effet substantiellement différentes, ne peuvent que difficilement expliquer aussi bien les mêmes résultats. Mon premier objectif sera donc de montrer que la distinction entre un système associationniste et un système basé sur des règles ne rend que très mal compte des résultats cités en son soutien. J'expliciterai ensuite la place que peut prendre cette distinction au sein de la théorie argumentative.

Réinterprétation des résultats

L'objectif de cette partie sera de montrer que la distinction proposée par Sloman ne peut pas réellement rendre compte des résultats qu'il appelle à sa défense. Il invoque plusieurs résultats expérimentaux en arguant qu'un type de réponse serait dicté par le système associationniste, et un autre par le système basé sur des règles. La stratégie qui va être utilisée est la suivante : pour chacun des résultats cités, j'essaierai de montrer que la réponse dont le système associationniste est censé être responsable ne peut être expliquée par ce système seul, et qu'il faut toujours une part de règles pour pouvoir en rendre compte. Il ne s'agit pas d'un argument définitif : on pourrait alors dire que bien que cette réponse nécessite des éléments d'association et de règles pour être expliquée, il s'agit tout de même de deux systèmes distincts. Cependant, dans la mesure où d'autres théories sont capables d'expliquer ces mêmes résultats de façon plus économique, cet argument devrait nous les faire préférer.

Etant donné que notre thème principal ici est le raisonnement, je passerai sur les données tirées de travaux sur la catégorisation. Pour ce qui est du raisonnement, Sloman introduit un critère, le Critère S, qui serait la marque du fonctionnement des

deux systèmes : « Un ensemble de données assez riche pour fournir un soutien substantiel à l'hypothèse de l'existence de deux systèmes de raisonnement existe. Ces données sont tirées de diverses tâches de raisonnement qui partagent toutes une unique et cruciale caractéristique. Elles satisfont toutes ce que je nomme le Critère S. Un problème de raisonnement satisfait au Critère S s'il cause chez les gens la croyance en deux réponses contradictoires » (p.11). Il cite ensuite une série de tâches qui, selon lui, satisfont ce critère, à commencer par le fameux problème 'Linda'.

Le problème de conjonction des probabilités

Ce problème fait partie d'une série de tâches mises au point par Tversky et Kahneman au début des années 80 afin d'étudier le raisonnement probabiliste et, plus particulièrement, la compréhension des conjonctions de probabilités – le théorème qui vient d'être cité en exemple de règle : $P(A) \geq P(A \& B)$. Afin de tester la robustesse de la compréhension et de l'utilisation de ce théorème par des participants naïfs, ils composèrent des tâches résolument trompeuses (Tversky & Kahneman, 1983). Elles impliquent la description d'une personne correspondant étroitement à un stéréotype répandu, l'exemple le plus fameux étant celui de Linda :

> Linda a 31 ans, elle est célibataire, très intelligente et n'hésite pas à donner son point de vue. Elle a fait des études de philosophie. En tant qu'étudiante, elle était très préoccupée par les problèmes de discrimination et de justice sociale, et elle a aussi participé à des manifestations anti-nucléaires.

Les participants doivent ensuite ordonner différents énoncés prétendant décrire Linda. Ces énoncés comprennent les deux suivants, faisant appel au stéréotype évoqué par la description :

> Linda est guichetière dans une banque
> Linda est féministe et guichetière dans une banque

Dans l'expérience originelle, plus de 80% des participants jugèrent que le second énoncé était plus probable que le premier, en violation du théorème qui vient d'être

évoqué (notons tout de même que ce jugement est, en quelque sorte, implicite : il n'est que reflété par le fait que le premier énoncé est placé plus haut sur la liste, les deux ne sont pas comparés directement). Les auteurs expliquent cet effet grâce à l'heuristique de représentativité : étant donné que la description est plus représentative de l'image d'une féministe que d'une guichetière de banque, le premier énoncé est favorisé. Sloman assimile cette heuristique à son système associationniste, ce qui ne semble pas tout à fait évident.

En effet, Tversky et Kahneman prennent soin de différencier leur notion de représentativité d'autres notions, plus élémentaires, de similarité :

> Cette relation diffère d'autres notions de proximité en ce qu'elle est distinctement directionnelle. Il est naturel de décrire un échantillon comme étant plus ou moins représentatif de sa population parente, ou une espèce (par exemple merle, pingouin) comme étant plus ou moins représentatif d'une catégorie superordonnée (par exemple les oiseaux). Il est malencontreux de décrire une population comme étant représentative d'un échantillon, ou une catégorie comme étant représentative d'une instanciation. (Tversky & Kahneman, 1983, p.296)

On voit donc qu'il ne s'agit pas d'un simple principe de similarité. De plus, il est très plausible d'imaginer qu'un ensemble de 'règles' (on dirait ici : de mécanismes inférentiels) soient nécessaires pour extraire le stéréotype à partir de la description. Si une notion de similarité joue un rôle, elle le joue à un très haut niveau conceptuel (le concept de 'féministe' n'est pas exactement ce qu'on pourrait appeler une primitive). Quoiqu'il en soit, on voit mal comment un simple réseau associationniste pourrait rendre compte d'une partie non triviale de la réponse donnée.

Raisonnement inductif

Le second exemple utilisé par Sloman est celui du raisonnement inductif, et de certains patterns d'erreurs qu'on peut y observer. Ce type de raisonnement est généralement étudié en demandant aux participants d'évaluer la force d'arguments

tels que celui qui suit (tiré de Sloman, 1993) :

> Tous les oiseaux ont une artère ulnaire
> Donc tous les merles ont une artère ulnaire

Un effet non normatif peut être observé si on compare la force attribuée à cet argument à celle qui est attribuée à l'argument suivant :

> Tous les oiseaux ont une artère ulnaire
> Donc tous les pingouins ont une artère ulnaire

Cet argument est jugé comme étant moins fort alors que, les pingouins étant des oiseaux, les deux arguments devraient être aussi forts – et maximalement forts. Un autre effet, lié, est celui d'inclusion, qui fait que l'argument suivant :

> Tous les merles ont une artère ulnaire
> Donc tous les oiseaux ont une artère ulnaire

Est jugé comme étant plus fort que celui-ci :

> Tous les merles ont une artère ulnaire
> Donc tous les pingouins ont une artère ulnaire

Alors que la conclusion du second est une conséquence de la conclusion du premier, et qu'il devrait donc être au moins aussi fort. L'explication que donne Sloman de ces erreurs est en termes de similarité : les merles sont plus similaires au prototype des oiseaux que les pingouins. On peut faire la même critique à cette explication que celle qui vient d'être faite dans le cas du problème Linda : le rôle que peut jouer la similarité est mal défini, et il s'insère dans un ensemble d'inférences qui dépassent les mécanismes associationnistes. La simple capacité de faire de tels jugements inductifs requiert des capacités inférentielles, à tout le moins un usage guidé de la similarité. Pour reprendre un argument classique (voir Goodman, 1955), tous les traits ne sont pas projetés de la même façon. Ainsi, si on remplaçait la conclusion de cet argument :

Tous les merles ont une artère ulnaire
Donc tous les oiseaux ont une artère ulnaire

par : « Donc tous les merles mécaniques ont une artère ulnaire », l'argument serait sûrement jugé comme étant maximalement faible. Or sur nombre d'aspects un merle mécanique est plus similaire à un merle qu'un oiseau prototypique. Il y a ici besoin de complexes mécanismes inférentiels pour pouvoir utiliser la similarité d'une façon un tant soit peu constructive. De plus, l'argument d'asymétrie évoqué dans le cas de la représentativité s'applique ici aussi : les catégories et les individus ne jouent pas des rôles similaires dans les mécanismes inférentiels (il semble par exemple que raisonner sur des individus soit considérablement plus aisé que de raisonner sur des catégories – voir Politzer & Mercier, In press).

Enfin, une théorie basée sur le principe de pertinence a été suggérée, qui rend compte non seulement de ces données, mais également des résultats de nouvelles expériences (Medin, Coley, Storms, & Hayes, 2003). Voici les conclusions des auteurs :

> Dans l'ensemble, les réponses démontrent, de façon robuste, l'importance des scénarios causaux et du renforcement des propriétés dans l'induction basée sur des catégories et fournissent un soutien pour le cadre pertinentiste. En plus des jugements portant sur la force des arguments, l'examen des justifications a révélé que, comme il était prédit, les propriétés des cibles et les relations causales étaient souvent mentionnées explicitement par les participants en expliquant les raisons de leur choix. (p.530)

Pour résumer, ces résultats montrent clairement que même dans des cas qui paraissent simples, les réponses des participants sont en fait guidées par de complexes mécanismes inférentiels : à la fois les scénarios causaux et le renforcement des propriétés qui sont mentionnées ne peuvent qu'être le résultat de tels processus. Il semble donc peu plausible d'attribuer à de simples phénomènes de similarité les résultats mentionnés par Sloman.

Biais de croyance dans les syllogismes

Le troisième phénomène mentionné par Sloman est celui du biais de croyance, tel qu'il a été observé par exemple dans le raisonnement syllogistique. Nous avons déjà vu en quoi le phénomène consistait dans la partie qui vient d'être consacrée à la théorie d'Evans. Sloman invoque ici la mémoire associative pour rendre compte des résultats dans le cadre de son système associationniste. Selon lui, une des réponses activerait en mémoire, par association, la croyance associée, et c'est cette activation qui expliquerait la réponse. Il est facile de voir à quel point cette explication est partielle, et ne rend compte que d'une partie relativement triviale du processus. Il est indéniable que pour que le biais de croyance existe, il faut que certaines de nos connaissances soient activées. Une telle activation est rendue inévitable par le fait même que les participants doivent lire et comprendre les énoncés.

Un tel processus de compréhension, et l'activation en mémoire des connaissances associées, sous-détermine cependant les réponses données. Tout d'abord, et comme Sloman lui-même le remarque, il y a toujours une interaction avec la logique (le fait qu'un argument soit valide ou non joue toujours un rôle, au-delà de la crédibilité de la conclusion). Mais surtout, on ne voit pas pourquoi la similarité, ou dans ce cas l'activation de connaissances associées en mémoire, devrait nécessairement faciliter l'acceptation d'une conclusion. Le fait que l'on accepte préférentiellement les énoncés qui sont conformes à nos croyances n'a rien d'évident, il ne découle pas de leur simple similarité. Dans le domaine perceptuel par exemple, ce n'est pas parce qu'une nouvelle information perçue est plus ou moins similaire à ce que nous avons en mémoire que nous devons plus ou moins l'accepter : dans tous les cas (ou presque), le contenu de la mémoire est simplement mis à jour. Pour expliquer ce qui se passe dans le cas du biais de croyance, il faut faire appel à des mécanismes spécialisés qui font qu'on évalue les informations communiquées à l'aune de nos croyances préalables – ce qui paraît tellement naturel qu'on est tenté d'oublier que ça n'est pas le cas général, et que ces opérations requièrent des hypothèses, et des mécanismes, additionnels.

Le raisonnement conditionnel et la tâche de sélection de Wason

Finalement, Sloman a recours à des résultats portant sur la tâche de sélection de Wason, le problème le plus utilisé pour étudier le raisonnement conditionnel[24]. Il tire parti en particulier d'un phénomène décrit plus haut : le biais d'appariement. Les effets de ce biais sont les suivants : les participants ont tendance à choisir les cartes qui sont mentionnées dans la règle, même si leur mention est niée (ainsi, si non-P est mentionné, c'est tout de même la carte P qui sera sélectionnée). Pour lui, il s'agit d'un phénomène associationniste car basé sur la similarité entre la carte mentionnée dans la règle et celle qui sera ensuite sélectionnée. Cependant, cet effet a depuis été réinterprété comme une conséquence du principe communicatif de pertinence (Sperber et al., 1995) (et ce de l'aveu même d'Evans, l'avocat originel du biais d'appariement, Evans & Over, 1996). L'idée est la suivante : lorsque nous interprétons des énoncés, certains éléments sont rendus pertinents durant le processus d'attribution d'une intention communicative au locuteur. Dans une des expériences conduites par Sperber et ses collègues, les participants doivent imaginer qu'ils sont des journalistes enquêtant sur une secte suspectée de pratiquer des inséminations artificielles sur des jeunes femmes vierges afin de produire des « mères-vierges ». Dans cette version du problème, la règle est « si une femme a un enfant, alors elle a eu des relations sexuelles », et les cartes à sélectionner sont « a un enfant », « n'a pas d'enfant », « a eu des relations sexuelles » et « n'a pas eu de relations sexuelles ». Dans ce contexte, l'énoncé de la règle rend particulièrement pertinentes non pas les cartes appariées à celles mentionnées dans la règle (« a un enfant » et « a eu des relations sexuelles »), mais bien les cartes désignant une potentielle « mère-vierge », c'est-à-dire « a un enfant » et « n'a pas eu de relations sexuelles ». Ce sont en effet ces cartes qui ont été sélectionnées par une majorité de participants (78%).

A travers cette expérience, et plusieurs autres exposées dans cet article, il apparaît clairement que le biais d'appariement n'est qu'un effet secondaire du principe de pertinence. Or les mécanismes par lesquels nous parvenons à attribuer une intention communicative à un locuteur sont complexes, hautement inférentiels, et si des processus associationistes y jouent sûrement un rôle, il n'est pas central. Ici

[24] Le raisonnement conditionnel étant le raisonnement qui implique des énoncés conditionnels du type 'si p, alors q'.

encore, l'explication en termes de similarité échoue donc à rendre compte des réponses qu'elle était censée expliquer.

Rôle des raisons dans les réponses 'basées sur des règles'

Nous venons de voir que, dans tous les cas cités en soutien de la théorie de Sloman, des mécanismes associationnistes ne peuvent pas expliquer les réponses qui sont typiquement attribuées au système 1. Des éléments vont maintenant être donnés en faveur de l'idée selon laquelle les réponses associées au système 2 sont non seulement basées sur des règles, mais sont plus précisément basées sur l'acceptation d'arguments, de raisons. Malheureusement, peu de recherches portent sur cette dimension des résultats : on admet généralement que les participants qui donnent la bonne réponse le font pour les bonnes raisons – qu'ils sont capables de rapporter – et que ceux qui ne l'ont pas donnée l'acceptent lorsqu'on leur explique. Ces deux éléments sont importants pour distinguer les bonnes réponses données, ou acceptées, sur la base du raisonnement stricto sensu (en tant que mécanisme métareprésentationnel) et celles qui sont données sur la base d'autres mécanismes inférentiels. Dans le cas du problème Linda par exemple, on pourrait imaginer que des participants possèdent un mécanisme psychologique dont le fonctionnement résulte dans le respect de la règle $P(A) \geq P(A \& B)$. Ils pourraient alors ne pas être capables de donner les raisons de leur réponse correcte. De même, on pourrait imaginer que des participants comprennent que la bonne réponse est en effet bonne dès qu'on leur présente, sans qu'il y ait besoin de leur fournir de raisons. Dans ces deux cas, il ne s'agirait pas de raisonnement à notre sens (bien que de telles opérations puissent tout à fait impliquer des 'règles', un point sur lequel je reviendrai dans la prochaine section).

Dans le cas des participants qui ne donnent pas spontanément la bonne réponse, il semble bien que s'ils finissent par l'accepter se soit parce qu'on leur présente de bonnes raisons de le faire[25]. En effet, dans toutes les expériences qui viennent d'être examinées, la bonne réponse ne peut guère ne pas avoir été considérée par les participants. Par exemple, dans le cas de la tâche de sélection de

[25] Ou par biais d'autorité, mais ce n'est guère pertinent ici.

Wason, on peut montrer que les participants prennent un peu de temps pour penser à chaque carte, ce qui aurait pu les amener à considérer la bonne réponse pour chaque carte (Evans, 1996; Roberts & Newton, 2002). De même, dans les problèmes d'induction, il est pour ainsi dire impossible de ne pas considérer toute l'échelle de force des arguments – ce n'est pas comme si la réponse logique était 'cachée'. On peut faire les mêmes remarques dans le cas du biais de croyance et du problème Linda. Ceci contraste avec les vrais cas d'insight, dans lesquels les participants n'entrevoient même pas la solution correcte (du moins ceux qui échouent bien sûr). Une fois qu'elle leur est communiquée, ils ressentent le fameux 'ahah !', ce qui n'implique pas (nécessairement) de raisonnement. Mais le cas des problèmes présentés ici est très différent : le 'ahah !', s'il survient, se situe au niveau de la compréhension des *raisons*, et de leur lien avec la réponse correcte. Par exemple, si on dit aux participants, dans la tâche de sélection de Wason standard, que la bonne réponse est P et non-Q, très peu comprendront immédiatement – après tout, ils viennent d'examiner et de refuser cette solution. Il faut leur expliquer, et ce n'est qu'alors qu'ils peuvent comprendre, et accepter, la bonne réponse (je ne connais pas de données montrant précisément cet effet, mais mon expérience d'avoir fait passer cette tâche de très nombreuses fois le confirme).

L'autre cas est celui des participants qui donnent spontanément la bonne réponse. Il faudrait alors savoir s'ils sont capables de la justifier adéquatement – ce que prédit la théorie argumentative. Mes données portant sur les problèmes du CRT (voir section 5.2.1) abondent dans cette direction, mais on pourrait arguer, à raison, que ces problèmes sont différents de ceux passés en revue ici en ce que la bonne solution ne se présente pas immédiatement aux participants : il faut nécessairement qu'ils raisonnent pour la trouver. Il paraît cependant très plausible que les participants donnant la réponse ont de bonnes raisons, et ce pour la raison suivante. Il est indéniable que la réponse intuitive est la première à laquelle l'immense majorité des participants accède. Ceux qui parviennent néanmoins à trouver la bonne réponse conservent généralement l'impression qu'elle est contre-intuitive ou, au moins, qu'une autre réponse est plus intuitive. Il serait dès lors surprenant qu'ils n'aient pas de raisons pour soutenir la bonne réponse, de raisons qui font qu'ils la préfèrent à l'autre. Imaginez que vous faisiez face à un problème. Une réponse se présente à vous, elle paraît très intuitive. A moins d'être totalement irrationnel, si vous donnez une autre réponse, c'est que vous avez des raisons pour le faire. Et il ne peut s'agir

du même type d'intuition, sinon c'est cette seconde réponse, la réponse correcte, qui serait la réponse intuitive.

Cette idée se trouve quelque peu soutenue par les résultats de raisonnement en groupe. Nous verrons dans la section 5.1.1 que pour le type de problème qui nous intéresse ici, le résultat des processus de raisonnement en groupe est le suivant : si un individu a la bonne réponse, alors il en convaincra les autres à coup sûr. Or, pour les raisons que j'ai évoquées plus haut, un tel changement d'opinion ne peut venir de la simple exposition à la réponse correcte : il doit venir de l'examen, et de l'acceptation, de raisons. Et ces raisons doivent bien être fournies par l'individu qui avait la réponse correcte dès le début (ou qui l'a trouvée durant le déroulement de la tâche). C'est ce que montrent clairement les transcriptions, dans lesquelles on peut observer ce processus d'échange d'arguments : les participants ne défendent pas la bonne raison en disant 'je sens que c'est ça', mais en essayant de démontrer aux autres qu'elle est, en effet, la bonne (voir Moshman & Geil, 1998; Trognon, 1993 et Mercier, résultats non publiés).

Intégrer la théorie de Sloman dans le cadre présent

Les arguments qui viennent d'être présentés pointent vers deux faits : d'une part des mécanismes inférentiels de nature non associationniste sont nécessaires pour expliquer les réponses typiques du système 1, et d'autre part des processus spécifiquement métareprésentationnels (et non simplement basés sur des règles) sont nécessaires pour rendre compte des réponses attribuées au système 2. La distinction proposée par Sloman ne correspond donc pas à celle que nous proposons – et les autres défenseurs de théories à processus duel seraient en accord au moins sur le premier point, à savoir que le système 1 n'est pas simplement associationniste (voir par exemple Evans & Elqayam, 2007, Stanovich, 2004, p.43). Faut-il pour autant ignorer cette distinction ?

Pas nécessairement. A mon sens, elle ne recouvre pas tant deux systèmes, ou deux ensembles de mécanismes cognitifs, que deux aspects essentiels du fonctionnement de tout système cognitif un tant soit peu complexe : d'un côté les mécanismes inférentiels, et de l'autre la régulation de ces mécanismes inférentiels. Dans leur définition la plus large, les mécanismes inférentiels sont des mécanismes

qui prennent un input, le traitent, et fournissent un output qui est informationnellement enrichi. Les mécanismes associationistes, quant à eux, joueraient plutôt un rôle de régulation. Il est nécessaire d'avoir des mécanismes permettant à l'organisme d'activer les bons mécanismes (inférentiels) dans les bons contextes. Bien qu'une partie des processus de régulation soit très probablement prise en charge par des mécanismes centralisés, une autre partie est sûrement distribuée, et peut en effet prendre la forme de mécanismes d'associations comme la règle de Hebb, selon laquelle deux éléments activés ensemble s'activeront plus facilement dans le futur si un s'est déjà activé.

On peut conclure en disant que ces deux aspects du fonctionnement cognitif sont également nécessaires. En l'absence de mécanismes inférentiels, un simple système associationiste n'est guère capable que d'accomplir les tâches les plus élémentaires – ce qui est bien montré par l'échec des théories béhavioristes à rendre compte du comportement d'animaux considérablement moins complexes cognitivement que les humains (voir Blaisdell, Sawa, Leising, & Waldmann, 2006; Leising, Wong, Waldmann, & Blaisdell, 2008; Murphy, Mondragon, & Murphy, 2008, pour des exemples de mécanismes basés sur des règles chez les rats). De même, sans mécanismes de régulation, dont certains prennent la forme d'associations, les mécanismes inférentiels ne pourraient pas être utilisés de façon appropriée.

La fonction du système basé sur des règles

Dans une courte partie précédant la conclusion de son article, Sloman s'interroge sur les fonctions possibles que peuvent desservir les deux systèmes qu'il décrit. Nous venons de voir que la dichotomie qu'il propose ne se situe probablement pas au même niveau d'analyse que celle des autres théories à processus duel, mais il est tout de même pertinent de s'intéresser aux réponses qu'il avance car elles pourraient tout aussi bien s'appliquer au système 2 tel qu'il est habituellement entendu.

La première suggestion est que les deux systèmes jouent des rôles complémentaires. Dans un cadre fonctionnaliste, cette suggestion est triviale dans la mesure où les deux systèmes sont différents. Ils ne peuvent dès lors desservir les

mêmes fonctions, et pourtant il faut bien qu'ils soient intégrés d'une certaine manière – ils sont alors quasiment forcément complémentaires. De plus, les exemples d'applications cités (mathématiques, loi, probabilités) sont tous des domaines culturels pour lesquels aucun des deux systèmes n'a pu évoluer.

La seconde idée est tirée de Freud, pour qui « inhiber ces processus primaires [l'équivalent ici du système associationniste] et ainsi rendre la gratification plus probable sur le long terme et le comportement plus acceptable socialement, est un processus de pensée secondaire [ici le système basé sur les règles] » (Sloman, 1996, p.18). Le problème de cette explication est que les mécanismes permettant de retarder la gratification et de rendre le comportement socialement acceptable sont partagés par de nombreuses espèces. La capacité à retarder la gratification est couramment étudiée chez les rats ou les oiseaux. Même si la capacité d'ajuster son comportement au monde social est plus rare, il est indéniable qu'un grand nombre de primates possède de telles facultés à un niveau très développé, et c'est également le cas de certains oiseaux (voir par exemple Clayton & Emery, 2007, pour le cas des corvidés). On peut dès lors adresser deux critiques à cette suggestion. D'une part, il est loin d'être évident que de telles capacités permettant de retarder la gratification et de se comporter adéquatement dans le monde social puissent réellement être définies comme un 'système' : il s'agirait plutôt d'un ensemble de capacités disparates ayant évoluées à des fins assez différentes. D'autre part, même si on pouvait effectivement penser à ces capacités comme à un système, il ne correspondrait pas à ce que les autres psychologues décrivent comme système 2, qui est beaucoup plus restreint et dont la possession est généralement vue comme étant propre aux humains. Il resterait donc à définir plus précisément la fonction de ce système plus restreint.

3.3 La théorie de Stanovich

Le domaine d'étude principal de Keith Stanovich est celui des différences interindividuelles dans les mesures d'intelligence générale, et c'est à ce sujet qu'il écrira un premier article, avec Richard West, introduisant sa conception de système 1 et de système 2 (Stanovich & West, 2000, voir également Stanovich, 1999). L'étude des différences interindividuelles n'est cependant que d'un intérêt marginal ici, et on se concentrera sur la version étendue de sa théorie, celle à laquelle il a consacré The

Robot's Rebellion en 2004. L'ordre d'exposition choisi par Stanovich sera repris, en ne passant en revue que les parties les plus pertinentes ici.

Les conflits cognitifs

Stanovich introduit le concept de processus duel d'une façon particulièrement frappante. Plutôt que de citer, comme cela est généralement fait, des exemples de comportements irrationnels dans le domaine du raisonnement (la tâche de sélection de Wason, le problème Linda, etc., sur lesquels il reviendra plus tard), il illustre le principe de « cerveau en guerre avec lui-même » par des cas de comportements socialement inadéquats. Il cite ainsi les quolibets dont sont victimes les personnes défigurées, ou le comportement des participants de l'expérience de Milgram. On peut douter cependant que ces exemples illustrent réellement la distinction entre les deux systèmes. Prenons le cas de l'étude de Milgram. Dans une série d'expériences, tellement fameuses qu'elles ont inspirées un film, I comme Icare, le psychologue Stanley Milgram demandait à des participants de donner des chocs électriques à d'autres personnes lorsqu'elles échouaient à remplir une tâche donnée (Milgram, 1974). Le prétexte était de tester l'effet de tels chocs sur les capacités d'apprentissage. Personne, en fait, ne recevait de choc, mais un complice mimait la surprise, puis la douleur et enfin l'agonie qu'une réelle personne aurait ressenti vu l'intensité des chocs prétendument délivrés (les participants ne faisaient qu'entendre le complice, sans le voir). Deux observations cruciales ont été dégagées de cette série d'expérience. Tout d'abord dans les 'bonnes' conditions, la majorité des participants obéit à l'expérimentateur et délivre le choc maximal, alors même qu'une menaçante tête de mort indique la dangerosité de cette intensité électrique. Il faut cependant que plusieurs conditions soient réunies : l'ascension doit être très graduelle, l'expérimentateur doit être très pressant, proche du sujet, en blouse blanche, prêt à assumer l'entière responsabilité, etc. La seconde observation, qui nous rassure quelque peu sur la nature humaine, est que les personnes qui croyaient avoir soumis d'autres individus à des chocs de cette intensité furent, généralement, traumatisées, et le mot n'est pas trop fort.

Pour Stanovich, il s'agit là d'un conflit entre le système 1 et le système 2. Il semble au contraire que les deux processus qui sont en conflit ici appartiennent au

système 1 : d'un côté l'empathie que nous ne pouvons que difficilement réprimer pour la personne que nous (croyons que nous) sommes en train de faire horriblement souffrir, et de l'autre l'envie d'obéir à l'expérimentateur qui nous presse de faire ce qu'il dit. La distinction entre système 1 et système 2 n'est pas une distinction entre un système 'gentil' et un système 'méchant'. Qui plus est, la réaction la plus viscérale, la plus 'système 1' est très certainement celle d'empathie : on le sait car la réaction des participants s'exprime physiologiquement (battements du cœur, transpiration, etc.). Si le raisonnement joue ici un rôle, c'est uniquement pour nous trouver des excuses nous permettant de continuer de faire un acte qui nous révulse. Cependant, pour que le raisonnement ait cette motivation même, il faut d'abord qu'un élan nous pousse à obéir, à faire ce que nous dit cette figure d'autorité – sinon il n'y aurait pas de conflit, pas d'actes, et rien à justifier. Si on peut louer l'intention de Stanovich de se distancier des exemples classiques, artificiels, d'irrationalité en prenant des exemples tirés des comportements sociaux qui ont beaucoup plus d'importance pour la plupart des gens, j'ai peur qu'il ne tombe dans le piège qui consiste à sur-interpréter tout conflit possible comme un conflit entre système 1 et système 2. Tous les exemples qu'il cite me semblent au contraire être des conflits entre différents mécanismes appartenant au système 1, et on pourrait tout aussi bien imaginer des conflits propres au système 2 (lorsque d'également bons arguments nous tirent dans des directions opposées).

Un nouveau nom pour le système 1 : TASS

Stanovich commence sa description du système 1 en insistant, à juste raison, sur le fait qu'il s'agit d'une assemblée de mécanismes, et non d'un simple système unifié. Il choisit donc d'y référer par l'acronyme TASS, pour 'The Autonomous Set of Systems' (l'ensemble des systèmes autonomes). Il reprend à son compte les grandes lignes des descriptions qui sont généralement données mais souhaite insister sur trois propriétés qui sont essentielles à ses yeux, et qui justifient l'importance du terme 'autonome'. La première est l'automaticité des processus TASS, la seconde est leur indépendance du contrôle central (du système analytique) et la troisième est le fait que les résultats fournis par les processus TASS entrent parfois en conflit avec un traitement parallèle effectué par le système analytique. Afin de mettre ces trois

propriétés en valeur, il reprend la liste des attributs prêtés aux systèmes modulaires par Fodor (Fodor, 1983). On peut donc en profiter pour reprendre ici cette même liste, en résumant le point de vue de Stanovich et en donnant la position qui découle du cadre adopté ici.

Stanovich commence par mettre de côté deux propriétés : celle d'avoir une architecture neurale fixe et celle d'être ontogénétiquement déterminé. Le cadre développé ici s'accorde sur ces deux points avec lui. Des processus peuvent parfaitement devenir modulaires tout en étant culturellement acquis (d'ailleurs, l'étude des expertises spécialisées débuta avec des exemples comme les échecs). Et pour ce qui est de l'architecture neurale fixe, bien qu'il s'agisse d'un bon élément en faveur de la modularité, il n'est nullement nécessaire – c'est ici le niveau computationnel, ou cognitif, qui doit rester prépondérant.

Si ce ne sont pas ces notions qui sont au cœur de l'idée de modularité, deux autres propriétés qu'il rejette, elles, le sont. Il s'agit des notions d'encapsulation informationnelle et d'impénétrabilité cognitive. La première signifie qu'un module n'a accès qu'à une base de données limitée pour traiter les informations qu'il reçoit, et la seconde que les processus centraux ne peuvent interférer avec le fonctionnement interne des modules. Stanovich mentionne que l'existence de ces propriétés est dure à démontrer empiriquement. Soit. Cependant, si on en prive les modules… il ne s'agit plus vraiment de modules. Le fait qu'un mécanisme ait un fonctionnement interne circonscrit, et relativement indépendant d'autres facteurs (en dehors des inputs) est ce qui définit la notion même de module dans tous les champs qui l'utilisent – la psychologie cognitive restant ici à l'écart car bloquée par la vieille définition de Fodor[26]. Il est d'autant plus étrange que Stanovich refuse d'attribuer ces propriétés aux modules qu'elles sont essentielles pour que les modules aient des propriétés que lui-même juge centrales. Il insiste en effet sur le fait que les modules sont rapides et que leur fonctionnement ne peut pas être altéré par les processus centraux. Mais s'ils sont rapides, c'est parce qu'ils ont une base de données limitée, et si leur fonctionnement n'est pas altéré par les processus centraux, c'est grâce à leur propriété d'impénétrabilité.

[26] Pour faire justice à Fodor, on doit mentionner d'une part que les critères n'étaient pas entendus comme obligatoires dans La Modularité de l'esprit (1983) et qu'ils se sont encore assouplis dans ses ouvrages plus récents (Fodor, 2001).

Une autre propriété que rejette Stanovich est celle de la spécificité de domaine. Selon lui, certains mécanismes appartenant au TASS sont 'domaine-généraux', tels les processus associationistes ou d'apprentissage implicite. Si on adopte la conception classique des domaines, comme la physique naïve ou la psychologie naïve, il est indéniable que certains processus échappent à cette classification. Il y a cependant deux moyens (au moins) d'étendre cette notion de domaine. D'une part, il peut s'agir de mécanismes métareprésentationnels. Dans ce cas, leur domaine est constitué par un ensemble de représentations, et ils s'attachent à certaines de leurs propriétés. Bien que cela puisse paraître surprenant, il n'y a aucune raison de rejeter un tel domaine ou d'en contester la spécificité. D'autre part, on peut penser que certains mécanismes de régulation exercent des fonctions très abstraites mais néanmoins très spécialisées. Ainsi, certaines régions des ganglions de la base se sont vu attribuer des rôles computationnels très précis dans l'apprentissage et la sélection d'action. Dans ce cas, on force vraiment nos intuitions sur ce qui peut constituer un domaine. On peut alors se demander s'il vaut mieux étendre le sens de domaine à ce type de computation, très abstraites, ou s'il faut trouver un nouveau terme. Quoiqu'il en soit, ces mécanismes sont très différents de ce qui est généralement entendu par processus central domaine-général : à nouveau, leur fonction est très spécifique et ils ne prêtent attention qu'à certaines propriétés des objets qu'ils traitent.

Nous avons vu qu'une des propriétés jugées centrales par Stanovich – l'indépendance du contrôle central – était en effet cruciale. Par contre, ce qui est pour lui la première caractéristique des modules, leur aspect 'inexorable' ('mandatory'), me semble être totalement impossible à respecter (voir par exemple p.52 : « L'essence des sous-processus TASS est qu'ils se déclenchent dès que le stimulus approprié est détecté, qu'ils ne peuvent pas être 'éteints' »). J'ai déjà défendu l'importance des processus de régulation à tous les niveaux cognitifs, et on peut citer des travaux montrant que mêmes les processus réputés 'automatiques' ont en fait besoin de ressources attentionnelles. Même en présence des bons stimuli, ceux qui sont censés déclencher inexorablement leur fonctionnement, des mécanismes d'aussi bas niveau que la lecture, ou même la vision des couleurs, ne sont pas activés (Joseph, Chun, & Nakayama, 1997; Rees, Frith, & Lavie, 1997, voir également Finkbeiner & Forster, 2008; Lien, Ruthruff, Cornett, Goodin, & Allen, 2008; Santangelo, Belardinelli, & Spence, 2007 et Pessoa, Kastner, & Ungerleider, 2002;

Pessoa, McKenna, Gutierrez, & Ungerleider, 2002, pour les émotions). Ils ne seraient donc pas automatiques si l'automaticité impliquait le caractère inexorable des opérations.

Il y a cependant ici une confusion possible entre le fait d'être automatique et d'être inexorable. Un mécanisme au fonctionnement inexorable est un mécanisme dont le fonctionnement ne peut jamais être stoppé : une fois le stimulus déclenchant présent, le mécanisme complète son fonctionnement. L'automaticité dépend elle plus précisément de l'absence de contrôle par les processus centraux. Si un mécanisme qui est inexorable est forcément automatique, il peut être automatique sans être inexorable : il existe d'autres mécanismes de régulation que les processus centraux. Dans le cadre de la compétition entre mécanismes cognitifs pour les ressources, on peut imaginer deux façons que pourraient avoir les mécanismes centraux d'influer sur les mécanismes périphériques. Une façon indirecte résulte de la simple activation de ces mécanismes centraux qui peut avoir pour conséquence d'épuiser les ressources, en privant ainsi certains mécanismes périphériques. Il s'agit là cependant plus de non-inexorabilité que de non-automaticité. Une vraie non-automaticité reviendrait, par exemple, à être capable de se concentrer sur ce que dit quelqu'un en étant capable de ne pas en comprendre le sens. Bien que cela puisse paraître surprenant, je pense qu'une telle chose est possible : peut-être pas en inhibant les mécanismes responsables du décodage linguistique, mais au moins en concentrant l'attention sur d'autres mécanismes – et ceci peut-être directement fait par les mécanismes centraux. Prenons le cas d'un opéra. Un expert en chant sera sûrement capable de concentrer toute son attention sur les propriétés esthétiques du chant, occultant tout à fait le décodage linguistique. C'est en fait ce que montrent les expériences mentionnées plus haut montrant la nécessité de ressources attentionnelles même pour des tâches de très bas niveau. Elles impliquent que les participants prêtent attention à une tâche particulièrement ardue tout en fixant un point donné. Si la tâche est bien calibrée, ce qui se trouve au point de fixation n'est quasiment pas traité. Or c'est bien ce qu'on considère comme les mécanismes centraux qui sont responsables de l'orientation de l'attention vers la tâche 'distractive'. On voit donc que les processus périphériques ne sont non seulement pas inexorables, mais qu'ils ne sont pas complètement automatiques non plus. C'est-à-dire que leur activation peut-être affectée par les opérations de mécanismes de plus haut niveau.

Pour finir, le troisième point sur lequel Stanovich insiste, et qui est lui indépendant des considérations fodoriennes sur les modules, est la possibilité de conflit entre les outputs du TASS et ceux du système analytique. Comme il a été mentionné plus haut, cette possibilité de conflit est indéniable, mais il n'y a pas de raisons de la distinguer des autres conflits possibles – soit entre mécanismes du TASS, soit entre mécanismes analytiques. En poussant cet argument à l'extrême, on peut rappeler qu'une des raisons pour l'émergence des mécanismes de régulation vient de la nécessité de gérer des conflits potentiels entre différents systèmes – et cela a probablement commencé avec l'apparition de la bilatéralité et de systèmes semi-indépendants gérant les deux côtés du corps (Prescott, Redgrave, & Gurney, 1999), on ne peut donc pas vraiment dire qu'il s'agisse d'un problème nouveau.

Le système analytique et le problème de l'homoncule

Stanovich ouvre la discussion sur le système analytique en mentionnant le danger de réduire ce système à un homoncule, un personnage fictif dans notre tête qui prendrait les décisions. Le problème de telles explications est qu'elles sont en fait vides : elles ne font que repousser d'un cran le besoin d'expliquer le comportement observé. L'objectif même des sciences cognitives est de réduire l'esprit, qui apparaît comme un tout intégré, en composants dont on peut décrire le fonctionnement de façon 'mécanique'. Si la réalisation d'un tel objectif semble très lointaine, il n'en reste pas moins que tout recours à des mécanismes qui sont aussi complexes que ce qu'ils sont censés expliquer ne peut pas contribuer à une meilleure compréhension du phénomène étudié. Stanovich, tout en acceptant cette critique épistémologique, défend cependant l'usage de la métaphore de l'homoncule pour parler du système analytique. Il a deux arguments. Le premier est que l'homoncule permet de parler du système analytique de façon plus simple – il serait vite lourd de mentionner à chaque fois la nécessité de décomposer cet homoncule. Il y a là à mon sens une confusion entre la nécessité de parler d'individus comme entités intégrées et celle de parler d'un système spécifique comme résumant presque entièrement ces individus. J'y reviendrai. Le second argument est tiré de Pinker (Pinker, 2000), qui dit, en somme, que bien que la métaphore de la société de l'esprit soit tout à fait valide, il faut bien que des mécanismes déterminent quel membre de cette société, quel mécanisme doit

s'exprimer à un moment donné, ce qui justifierait de parler en termes d'homoncule. Même en acceptant l'argument de Pinker (sous une forme quelque peu modifiée peut-être), il faut noter qu'il ne peut s'appliquer au problème de Stanovich : Pinker, en effet, se situe dans un cadre massivement modulariste, et Stanovich non. Or l'argument de Pinker ne fonctionne que dans ce cadre. Pour le comprendre, il faut revenir aux sources et à l'idée de mécanismes centraux non modulaires chez Fodor.

Fodor oppose aux systèmes périphériques, modulaires, des systèmes centraux non modulaires desservant les fonctions les plus 'hautes' (mise à jour des croyances, raisonnement scientifique, etc.). Il s'agit pour lui d'un constat d'échec : étant donné qu'on ne peut pas, pense-t-il, décomposer ces systèmes, on ne pourra jamais comprendre leur fonctionnement. Même lorsqu'elle fait face à des systèmes infiniment plus simples que ces systèmes centraux, la science a besoin de décomposer pour comprendre. On voit donc mal comment elle pourrait comprendre le fonctionnement d'un système aussi complexe, le plus complexe qui soit peut-être, sans le décomposer. Etrangement, les partisans des théories à processus duel, qui assimilent souvent leur système 2 à ces processus centraux, ignorent cette conclusion et continuent d'essayer d'en comprendre le fonctionnement comme si de rien n'était. En attendant que quelqu'un montre comment on peut comprendre un tel système sans le décomposer, la seule solution viable est d'abandonner cette hypothèse d'amodularité, en d'autres termes, d'adopter le cadre de la modularité massive (voir Sperber, 1994). C'est ce qu'ont fait plusieurs psychologues évolutionnistes dont Pinker se réclame (Pinker, 2000).

Le problème de l'homoncule est différent pour Stanovich – qui se place dans la perspective de systèmes centraux amodulaires – et pour Pinker (ou d'autres) qui acceptent l'hypothèse de modularité massive. Dans le cas de Stanovich, il s'agit d'un réel problème : dire que les systèmes centraux prennent des décisions, par exemple, ne fait que repousser toute explication. Dans le cas de Pinker, le problème est différent et devient le suivant : peut-on toujours parler d'individu, en tant que mécanisme intégré, alors qu'on postule que l'esprit est une somme de mécanismes fonctionnant en autonomie partielle ? Certains psychologues évolutionnistes répondent non à cette question (Kurzban & Aktipis, 2007, par exemple), et c'est à ceux-ci que s'oppose Pinker. Et à raison : comme il le dit, les différents mécanismes qui composent notre esprit doivent bien être intégrés à un certain niveau. C'est là le rôle des mécanismes de régulation qui ont été mentionnés à de nombreuses reprises.

Il n'est donc pas correct d'utiliser l'argument de Pinker en soutien d'une vision homonculaire des systèmes centraux – au contraire, Pinker et les autres partisans de la modularité massive se battent justement contre une telle perspective.

Le système analytique comme 'machine virtuelle'

Dans la seconde partie consacrée au système analytique, Stanovich reprend et élabore sur la métaphore proposée par Dennett de la conscience comme une 'machine virtuelle' (Dennett, 1991). Dans La Conscience expliquée, Dennett défend l'idée que les processus conscients (le système 2) sont le produit d'une machine virtuelle sérielle qui recrute les mécanismes neuronaux fonctionnant en parallèle (le système 1). Dans le domaine de l'informatique, de telles 'machines virtuelles' sont par exemple utilisées pour émuler le fonctionnement d'un système d'exploitation Windows sur un Macintosh. Sous cette description assez vague, il me semble qu'on retrouve l'idée que les processus analytiques sont métareprésentationnels : ils peuvent recruter les processus intuitifs en activant les représentations sous jacentes. On retrouve là l'idée que, comme le dit Stanovich : « les outputs domaine-spécifique des modules du TASS puissent être recrutés pour servir des fins plus générales, accroissant ainsi la flexibilité du système ». Il est difficile de contester cette interprétation – relativement floue, avouons le – des faits psychologiques. Le problème se situe plus, comme c'était déjà le cas pour Evans, au niveau des explications ultimes. Ayant reconnu que le système analytique nous donne une certaine flexibilité comportementale, il semble difficile de ne pas y voir là l'une des ses fonctions principales, ce qui a déjà été contesté dans la partie portant sur les explications évolutionnistes d'Evans.

Un autre aspect important de cette 'machine virtuelle' serait son lien avec le langage. Selon Stanovich : « Virtuellement tous les chercheurs en sciences cognitives s'accordent sur le fait que le système analytique seul est réceptif aux inputs linguistiques, qu'ils soient d'origine interne ou externe » (p.48). La vérité de cet énoncé présuppose une conception extrêmement large du système analytique. Par exemple, même s'ils n'ont que des effets limités, des stimuli langagiers présentés en deçà du seuil de la conscience peuvent avoir des conséquences comportementales liées à leur signification (voir la section 4.1). D'autre part, et plus généralement, cette

perspective me semblent être très dépendante d'une vision du langage qui ne prend en compte que la syntaxe et la sémantique. Si on prend on compte l'étage pragmatique, alors la distinction entre langage et autres formes de communication se brouille. Or il me semble être assez clair que des formes non verbales (ou plutôt, les composants non linguistique de la communication, même verbale) activent assez directement des systèmes appartenant au TASS : un cri de détresse, les pleurs d'un nourrisson, une intonation triste ou joyeuse peuvent tous activer des émotions sans, apparemment, passer par un système analytique.

Dans le cadre de la théorie argumentative, le lien entre langage et raisonnement se trouve plutôt au niveau des causes ultimes. Si le raisonnement a évolué pour l'argumentation, il nécessite des capacités communicatives. Bien que des mécanismes aussi complexes que le langage ne soient pas forcément nécessaires, la complexité des mécanismes de communication est le reflet de son importance (si la communication n'est pas importante, nul besoin de mécanismes aussi sophistiqués) et c'est cette importance même qui à son tour requiert des mécanismes tels que l'argumentation pour que les individus puissent en tirer des bénéfices maximum. Il y a donc de bonnes raisons de penser que le raisonnement a évolué, à tout le moins c'est perfectionné, en même temps que se développaient des formes de communication plus sophistiquées. De plus, une fois que le langage est devenu le medium de communication par excellence, c'est lui que les gens utiliseront principalement pour argumenter, liant ainsi langage et raisonnement pour ce qui est de leur utilisation – ce qui n'implique nullement un lien au niveau des mécanismes sous-jacents (ce n'est pas parce que différentes parties du système digestif tendent à être co-activées qu'elles doivent avoir des parties en commun).

Le système analytique et la pensée hypothétique

De même qu'Evans, Stanovich lie système analytique et pensée hypothétique : « Une des fonctions du système analytique est de soutenir la pensée hypothétique » (p.50)[27]. Il développe d'ailleurs plus les liens entre ce système

[27] Une position qui est cependant moins forte que celle d'Evans, pour qui le système analytique semble se résumer à la pensée hypothétique, ou tout au moins n'opérer que sur ce type de représentations.

hypothétique tel qu'il est habituellement étudié dans les domaines du raisonnement et de la prise de décision et les recherches portant sur les autres formes de 'découplage', en particulier la théorie de l'esprit. Il mentionne également les métareprésentations, qui rendraient possible le « regard d'autocritique qui est un aspect unique de la cognition humaine » (p.51). Il semble régner une certaine confusion entre les notions de pensée hypothétique, de découplage, et de métareprésentation. Dans la partie sur Evans, j'ai défendu l'idée qu'il était préférable de se limiter aux termes de représentation et de métareprésentation. Dans ce cadre, la notion de métareprésentation est très large : il s'agit, rappelons-le, de représentations de représentations. Malheureusement, Stanovich la limite aux « représentations de nos propres représentations » (p.51) – il en exclut donc des capacités comme la mentalisation, c'est à dire la capacité de représenter les états mentaux d'autrui. Pour cette raison, il limite les mécanismes métareprésentationnels à ce qui est souvent appelé métacognition, ou capacités métacognitives, à savoir la capacité d'évaluer la façon dont nous pensons. S'il est en effet pertinent de séparer différents types de mécanismes métareprésentationnels – ceux de la mentalisation et ceux du raisonnement par exemple – il s'agit dans tous les cas de métareprésentations.

Les interactions : inhibition du TASS par le système analytique

Bien que Stanovich concède le caractère adaptatif des processus TASS, il insiste sur le fait qu'ils donnent tout de même parfois des réponses inappropriées (particulièrement dans l'environnement moderne), et donc sur la nécessité dans laquelle se trouve alors le système analytique de les supplanter. De plus, même si « cette fonction de supplantation [override] n'est nécessaire que dans un nombre limité de situations de traitement d'information, il peut s'agir de situations très importantes » (p.62). Le problème est que cette fonction n'est nécessaire que parce que les processus du TASS sont conçus comme ayant un fonctionnement automatique et inexorable. Elle devient superflue si on prend en compte le fait que ces processus ne sont en fait pas inexorables, et que leur activation est régie par des processus régulateurs complexes.

Pour défendre la nécessite du rôle de supplantation du système analytique, Stanovich s'inspire du modèle (lui aussi à processus duel) développé par Pollock

dans le domaine de l'intelligence artificielle. Comme le dit Stanovich : « Pollock met l'accent sur le fait qu'une fonction importante des traitements analytiques est de supplanter le TASS quand un module rapide et inflexible [l'équivalent du TASS chez Pollock] s'est déclenché dans un environnement dans lequel sa réponse rigide n'est pas bien adaptée – un environnement modifié avec lequel le module rapide et inflexible ne peut pas traiter 'en direct' car sa vitesse dépend de la rigidité avec laquelle il réagit ». Il me semble que Stanovich se méprend sur la cause de la rapidité de traitement de certains modules. Elle n'est pas seulement due au fait qu'ils se déclenchent automatiquement, mais également au fait qu'ils n'ont accès qu'à une base de données limitée. On pourrait très bien imaginer un module se déclenchant dès qu'il rencontre un certain stimulus, mais qui doive chercher dans une base de données énorme afin de pouvoir le traiter – cela pourrait le ralentir considérablement. Les mécanismes de régulation qui font que mêmes les mécanismes les plus proches des réflexes ne sont pas inexorables peuvent être rapides pour deux raisons.

La première est que les mécanismes de régulation centralisés – comme certaines structures des ganglions de la base – ont été peaufinés par l'évolution pour prendre des décisions extrêmement rapidement, et ce au moyen de zones spécialisées effectuant des computations bien précises. La seconde est qu'une bonne part de la régulation intervient en aval de l'exposition même aux stimuli. Par exemple, un rat dans un espace ouvert aura un seuil de détection des prédateurs plus bas que s'il est dans la sécurité du fond d'une sombre niche. Dans ce cas, le mécanisme réagira aussi rapidement une fois déclenché, mais il n'empêche que ce déclenchement même sera modulé par la pertinence du mécanisme dans l'environnement présent – ce qui correspond exactement à ce que Stanovich, à la suite de Pollock, attribue au système analytique. Voici un exemple révélateur. Le mouvement de défense des écrevisses, qui pourrait passer pour un réflexe typique, desservi par un type spécial de fibre, est en fait modulé par l'environnement physique (est-ce qu'il y a assez de place pour fuir ?), par la prise en compte d'autres besoins (est-ce que je dois quitter une source de nourriture ?) et même par l'histoire sociale de l'individu (un dominant ne s'enfuit pas ! – enfin… plus difficilement) (D. H. Edwards, Heitler, & Krasne, 1999). Il ne s'agit ici nullement d'un embryon de système analytique qu'on trouverait chez les écrevisses, mais bien des mécanismes de régulation qui sont nécessaires à tous les systèmes cognitifs un tant soit peu complexes.

Sur la fonction du système analytique

Après avoir passé en revue les grandes lignes du fonctionnement des processus TASS et du système analytique, Stanovich s'étend sur leurs fonctions respectives. Il propose à ce propos un cadre original, inspiré des débats sur le niveau de sélection en biologie. Depuis la parution du Gène égoïste de Dawkins (1990), ou plutôt des travaux qui l'ont inspiré, les biologistes s'interrogent sur le niveau de sélection qui est pertinent pour l'évolution : celui du gène ? De l'individu ? Du groupe ? La version la plus classique de ce débat (pour ce qui est de l'évolution de l'homme), se situe entre le niveau de l'individu et celui du groupe pour expliquer l'apparition de l'altruisme. Stanovich se place lui à un niveau différent : « La structure des objectifs du TASS a été formée par l'évolution pour suivre de près les augmentations des probabilités de reproduction des gènes. Le système analytique est essentiellement un système de contrôle focalisé sur les intérêts de la personne en entier. C'est le mécanisme de maximisation principal de la satisfaction des objectifs *personnels* d'un individu » (p.64). Pour lui donc, il y aurait un conflit entre le niveau des gènes et celui de l'individu. De quoi il pourrait s'agir n'est pas très clair. Une possibilité, que l'on peut exclure immédiatement, est celle de conflits intragénomiques : il s'agit de gènes qui tentent de se reproduire au détriment d'autres gènes – par exemple en biaisant le processus de la méiose (voir Burt & Trivers, 2006, pour une revue). Il s'agit là de conflits entre d'un côté un gène (ou ensemble de gènes) isolé et les autres gènes. Stanovich se place à un autre niveau, celui de l'ensemble des gènes versus le niveau de l'individu.

On peut alors imaginer au moins deux interprétations : dans la première, le niveau de l'individu serait celui de la satisfaction personnelle, alors que dans la seconde il s'agirait uniquement de sa survie. Dans les deux cas, ce niveau s'opposerait à celui des gènes qui eux ne viseraient que leur propre reproduction. On peut éliminer facilement celui de la satisfaction personnelle : les mécanismes déterminant ce qui nous rend heureux, ce qui nous fait plaisir, ont en effet été façonnés par l'évolution depuis de centaines de millions d'années, et ils correspondraient plutôt à ce que Stanovich vise par le niveau des gènes (ils ont été façonnés pour maximiser notre reproduction). Reste alors le niveau de la survie. C'est ce que semble indiquer Stanovich en donnant l'exemple suivant de conflit entre

gènes et individu : « Les objectifs [qui reflètent uniquement les intérêts des réplicateurs] sont ceux qui sacrifient le véhicule aux intérêts des réplicateurs – ceux qui amènent une abeille à se sacrifier pour la reine qui lui est génétiquement apparentée » (p.65). Mais on est alors confrontés à un réel dilemme : l'évolution d'un tel système, qui favorise la survie par rapport à la reproduction est… impossible. Imaginons la situation suivante. Des individus se retrouvent face à un choix. S'ils choisissent l'option A ils survivent mais seront stériles[28]. S'ils prennent l'option B, ils meurent mais avant de mourir peuvent laisser un descendant. Ce sont les individus qui prennent l'option B qui seront sélectionnés. Toujours.

Stanovich pourrait répondre qu'il s'agit là d'une vision simpliste, et que l'existence de tels systèmes devient possible si en prend en compte la division du travail cognitif. Admettant que survie et reproduction sont des problèmes décomposables, certains systèmes pourraient être chargés principalement de la première, et d'autres de la seconde. Stanovich, semble-t-il, ne dit pas autre chose lorsqu'il propose que les systèmes TASS visent surtout la reproduction alors que le système analytique viserait surtout la survie. Cet argument est plausible. Ce qui ne l'est pas, cependant, c'est que le système qui vise la survie l'emporte sur celui qui vise la reproduction en cas de conflit. Ou plutôt, étant donné que des mécanismes doivent servir d'arbitre dans ces cas là, il est impossible qu'ils aient évolué afin de favoriser la survie au détriment de la reproduction. Postuler des systèmes distincts ne fait donc que repousser le problème. Pour Stanovich au contraire, le système analytique serait juge et partie : non seulement il serait responsable de prendre des décisions favorisant l'individu (ou plutôt sa survie), mais il serait également responsable de faire que ce soient ces mécanismes qui l'emportent en cas de conflit. Si des mécanismes existent bien qui visent la survie et non la reproduction, je ne vois pas comment des mécanismes pourraient évoluer qui favorisent la première au dépend de la seconde.

Notons un autre problème avec la vision de Stanovich. Pour lui, « de nombreux objectifs instanciés dans ce système [le TASS] furent acquis de façon non réflexive – ils n'ont pas subit d'évaluation en termes de leur utilité pour les intérêts de la *personne*. En fait ils ont été évalués, mais par un différent ensemble de critères :

[28] En assumant que l'individu qui survit ne pourrait pas augmenter son aptitude globale en aidant des apparentés à se reproduire.

est-ce qu'ils augmentaient la longévité et la fécondité des réplicateurs dans le passé évolutionniste » (p.64). En d'autres termes, l'évolution aurait 'oublié' les mécanismes du TASS, et plutôt que de les corriger, elle aurait mis au point un nouveau système, le système analytique, qui serait vraiment au service des intérêts de la personne. Cet argument n'est pas plausible. Si les pressions de sélection ont été assez fortes, et que le temps nécessaire était disponible pour la création d'un tout nouveau système cognitif, il n'y a aucune raison de penser que les mécanismes précédents n'aient pas pu évoluer pour s'adapter à ces nouvelles pressions. Cela va à l'encontre de ce qu'on sait du fonctionnement de l'évolution, qui se cantonne le plus souvent à des 'bricolages', pour reprendre le mot de François Jacob, et n'est pas dispendieuse au point de créer un nouveau mécanisme alors que des ajustements de mécanismes préexistants pourraient suffire. Stanovich pourrait alors arguer que le propre des systèmes TASS est justement d'être inflexible, mais ce n'est pas cohérent avec ce qu'on sait de ces mécanismes et même des mécanismes évolués en général, pour lesquels l'évolvabilité est toujours un critère important (Kirschner & Gerhart, 1998; Pigliucci, 2008; Wagner & Altenberg, 1996).

Ce problème est bien illustré par un exemple que donne Stanovich d'inflexibilité des processus TASS, et par là de la nécessité du système analytique :

> Dans notre environnement présent, cependant, bien que le TASS puisse tout à fait signaler automatiquement à un mâle de s'accoupler avec une femelle entr'aperçue ("fonce !"), le système analytique détecte correctement que l'homme vit dans une société technologique complexe au vingt et unième siècle et que des considérations comme une épouse, des enfants, un travail et position sociale demandent que ce "fonce !" du TASS soit supplanté (p.63).

Il y a bien longtemps que les animaux vivant en groupe ne se permettent pas ce genre d'incartades ! Dès qu'une hiérarchie s'établit chez les mâles, une des prérogatives des dominants est d'avoir un accès sexuel favorisé aux femelles – chez les poules aussi bien que chez les chimpanzés (chez les poules par exemple, « 3 % des mâles réalisent 87 % des copulations », Campan & Scapani, 2002, p.554). Pour que les individus ne soient pas en guerre constante, il faut bien que les subordonnés inhibent leurs pulsions. On voit bien qu'il est tout à fait possible de réguler des comportements sans système analytique (en faisant l'hypothèse que Stanovich, et les

autres partisans des théories à processus duel, refuseraient de prêter un système analytique aux coqs – ce qui semble raisonnable car ils s'accordent sur le fait qu'il n'a évolué que très récemment). Les explications évolutionnistes avancées par Stanovich ne sont donc, dans l'ensemble, guère crédibles.

3.4 Conclusion sur les théories à processus duel

De nombreux aspects des différentes théories qui viennent d'être présentées ont été critiqués, mais il ne serait cependant pas approprié de les rejeter entièrement. Elles offrent chacune des éléments permettant de mieux comprendre le fonctionnement du raisonnement – en particulier celle d'Evans, qui propose le modèle le plus précis sur ce point. Le domaine pour lequel elles sont le plus faible est probablement celui touchant à la fonction du raisonnement, aux explications évolutionnistes. Nous avons vu qu'alors ces différentes théories s'avèrent être très vagues (c'est le cas d'Evans et de Sloman) ou reposer sur une vision des mécanismes du système 1, et de leur interaction avec ceux du système 2, qui est difficile à réconcilier avec de nombreuses données empiriques.

Il est malgré tout important de conserver à l'esprit le fait assez remarquable que toutes ces théories convergent vers l'idée qu'une partition doit être faite entre deux grands types de processus. Ceci est d'autant plus frappant que des chercheurs de nombreuses autres disciplines sont arrivés à des conclusions similaires. Si les dichotomies utilisées dans les champs de l'attention (Posner & Snyder, 1975) et de la mémoire (Schacter, 1987) ne correspondent peut-être pas exactement à celle qui est utilisée dans le domaine du raisonnement, la division pensée par les chercheurs en psychologie sociale (Chaiken & Trope, 1999; Gawronski & Bodenhausen, 2006; Wilson, Lindsey, & Schooler, 2000) et en prise de décision (Kahneman & Frederick, 2002, 2005; Thaler & Sunstein, 2007) est par contre très proche (elle s'en inspire même directement dans le dernier cas). Ceci nous permettra, dans la seconde partie, de faire appel à des résultats venant de toutes ces disciplines afin de soutenir la théorie argumentative du raisonnement.

Soutien empirique

Dans cette partie, je vais montrer que de nombreuses données soutiennent la théorie argumentative du raisonnement. Le premier chapitre est consacré non pas au raisonnement lui-même, mais à un de ses précurseurs, un autre mécanisme de vigilance épistémique avec lequel il interagit : la vérification de cohérence. Le reste sera consacré à étayer différentes prédictions de la théorie argumentative. La première, la plus directe, est que le raisonnement devrait être efficace pour produire et examiner des arguments, et ce d'autant plus qu'il est utilisé dans un contexte normal – une discussion entre des personnes qui tentent de résoudre un désaccord par le biais d'arguments. Le chapitre deux s'ouvrira donc par une revue des travaux pertinents concernant le raisonnement en groupe. Puis les résultats d'autres tâches qui mettent les participants dans une situation d'argumentation seront passés en revue. Je m'intéresserai ensuite, dans le chapitre trois, aux résultats, venus principalement de la psychologie du raisonnement, montrant que les gens sont biaisés, mais pas de n'importe quelle façon : ils ont un fort biais de confirmation vis-à-vis des idées cohérentes avec leurs croyances et un biais d'infirmation pour les autres idées. Si ces biais s'accordent parfaitement avec le fonctionnement du raisonnement prédit par la théorie argumentative, ils sont des épines dans les pieds de ses concurrents. Dans le chapitre quatre seront passées en revue les études sur le raisonnement motivé et ses apparentés. Leurs résultats montrent que les gens ne raisonnent pas 'objectivement', mais partent d'une conclusion, avantageuse pour eux, qu'ils essaient de défendre, et qu'ils ne la rejettent que s'ils échouent à trouver des justifications plausibles. Dans le chapitre cinq, un courant important dans la psychologie de la prise de décision sera examiné : celui du choix basé sur des raisons. Nous verrons que de très nombreuses démonstrations d'« irrationalité » dans ce domaine peuvent s'expliquer facilement en faisant l'hypothèse que le raisonnement n'a pas pour fonction de nous faire prendre de meilleures décisions, mais de nous faire prendre des décisions qu'on peut justifier. Ceci sera confirmé par une analyse des résultats montrant que, dans de nombreuses tâches, l'utilisation du raisonnement ne permet pas une amélioration des performances – contrairement à ce que devraient prédire les théories classiques. Mais avant tout cela, quelques prédictions générales faites par la théorie argumentative vont être présentées,

prédictions qui pourront servir de fil directeur pour cet exposé des résultats empiriques.

Prédictions sur les contextes d'activation du raisonnement

La théorie argumentative permet de faire des prédictions à la fois sur les circonstances qui devraient avoir tendance à activer le raisonnement, et sur la façon dont ces contextes devraient influencer son fonctionnement. Selon la théorie, le contexte le plus naturel d'activation du raisonnement est le rejet d'une information par la personne à qui on souhaitait la communiquer. Ce rejet peut s'observer de plusieurs façons : il peut être communiqué explicitement (la personne nous dit qu'elle n'est pas d'accord, qu'elle ne veut pas faire ce qu'on lui demande), communiqué implicitement (la personne nous laisse entendre son désaccord) ou simplement observé (la personne se comporte d'une manière incohérente avec l'acceptation de ce que nous avons communiqué). Bien que comprendre que ce que nous avons voulu communiquer a été rejeté puisse impliquer des mécanismes différents selon les cas, le résultat final est sensiblement identique : ce rejet nous fournit une motivation pour essayer de convaincre l'interlocuteur (à moins que nous ne préférions avoir recours à des stratégies différentes qui ne reposent pas sur le raisonnement comme lui faire peur, lui rappeler qui est le chef, etc.). Notons tout de même que dans un cadre coopératif optimal, le désaccord devrait être communiqué explicitement ou, s'il l'est implicitement, de telle façon que l'on soit sûr qu'il soit compris. De plus, un désaccord explicite s'assortira souvent d'une justification qui, en précisant les raisons du désaccord, rendra la recherche d'arguments plus aisée.

Dans d'autres cas, toujours de nature argumentative au sens classique, le désaccord sera anticipé. Il arrive en effet que nous voulions communiquer quelque chose à une personne mais que nous sachions qu'elle ne l'acceptera pas immédiatement. Nous pouvons alors anticiper sa réaction en préparant des arguments que nous serons soit prêt à présenter en cas de désaccord manifeste, soit que nous pourrons utiliser de façon préventive en les énonçant avant même de s'avancer sur la conclusion que nous voulions initialement communiquer.

Enfin, le raisonnement sera également activé lorsque nous devons évaluer un argument qui nous est présenté. Ceci aura généralement pour effet d'augmenter les

chances que la conclusion soit acceptée – dans la mesure où l'argument est bon. Pour cette raison, les locuteurs ont intérêt à souligner les liens logiques entre leurs propositions, à souligner leurs raisonnements. Ils peuvent pour cela utiliser nombre de termes logiques (donc, si… alors) ou paralogiques (puisque, car). Ces termes auront pour effet d'orienter l'attention de l'interlocuteur, de le faire considérer certaines propositions en tant qu'arguments soutenant une conclusion et, partant, d'augmenter les chances que cette conclusion soit acceptée. Dans de nombreux cas, la phase d'évaluation 'pure', qui consiste à se représenter une proposition en tant que raison, sera suivie d'une phase de production : à moins que l'argument ne soit considéré comme définitif, l'interlocuteur à qui il était destiné cherchera des contre-arguments. Cette recherche de contre-argument peut, à son tour, avoir un rôle évaluatif indirect : si elle est trop difficile, infructueuse, elle peut conduire la personne à réévaluer sa position, ne serait-ce que sa position publique si l'échange à des spectateurs.

Cependant, il est indubitable que le raisonnement est souvent utilisé dans des tâches de prise de décision qui semblent ne rien avoir d'argumentatif. Je vais défendre l'idée que dans ces situations, ce sont bien les mêmes capacités argumentatives qui sont utilisées, mais qu'elles le sont alors pour montrer aux autres que la décision que nous avons prise ou que nous allons prendre est une bonne décision (c'est-à-dire que nous sommes compétents et bienveillants en prenant cette décision). D'une manière analogue aux cas argumentatifs 'normaux', de telles situations surviennent lorsque nos décisions sont susceptibles d'être mises en question ou jugées négativement par d'autres : nous voudrons alors trouver des arguments afin de leur montrer qu'il s'agissait en fait d'une décision bonne, à tout le moins justifiable. Mais on retrouvera ici les cas de jugements anticipés : si nous craignons qu'une de nos décisions soit mal perçue par les autres, nous pouvons commencer à préparer des arguments la défendant, afin de pouvoir les utiliser si la décision est effectivement remise en cause ou même en les joignant spontanément à notre décision de façon à anticiper un jugement négatif.

Quels facteurs font que nos décisions sont susceptibles d'être mal jugées par d'autres personnes ? Tout d'abord, plus les décisions sont publiques – plus elles sont facilement observées par d'autres – et plus il y a un risque qu'elles soient évaluées. Cela ne veut cependant pas dire qu'on puisse s'affranchir aussi facilement de la possibilité d'être jugé lorsqu'on est seul, et ce pour au moins deux raisons. La

première est que nombre de nos décisions vont laisser des traces qui pourraient être observées par d'autres (vous décidez de vider la boîte de chocolats en l'absence de votre compagne ou compagnon). La seconde est que les autres sont tellement omniprésents dans notre vie qu'il est tout à fait raisonnable d'agir, en première approximation, comme si tous nos actes étaient publics, ou avaient une chance de l'être ou de le devenir. Mais cela ne veut pas dire que le raisonnement sera dès lors utilisé uniformément : il y aura des variations selon que la décision est plus ou moins publique, ainsi que selon la taille et la composition du public. Certaines décisions sont clairement publiques (vous participez à un jeu télévisé), d'autres vont presque certainement rester privées (notons tout de même que le risque que vous révéliez vous-mêmes votre action, même en l'absence de tout indice matériel, n'est presque jamais nul). De plus, certain publics nous importent beaucoup plus que d'autres – nos familles, nos amis, nos supérieurs hiérarchiques. Enfin, ces effets du public ne sont pas uniquement quantitatif, mais bien qualitatifs : une décision, ou une justification, qui siérait à un public pourrait fortement déplaire à un autre.

Il s'agit là des facteurs externes, mais ils ne peuvent être les seuls à rentrer en jeu : nous ne nous préparons pas à justifier toutes les décisions que nous prenons – heureusement ! Comment donc savoir quelles décisions risquent d'être jugées comme étant mauvaises ? Dans certains cas, nous savons clairement que nous faisons quelque chose qui déplaira à quelqu'un : il peut alors être bon, si nous décidons de nous engager malgré tout dans cette voie, d'avoir des arguments prêts. Mais il me semble que la majorité des cas correspondent à des incertitudes : nous ne sommes pas sûrs de nous, et il y a donc des chances que ce que nous faisons soit mal jugé. Deux types de circonstances peuvent entraîner une telle incertitude. Dans le premier, nous avons des intuitions assez fortes, mais contradictoires : aucune ne l'emporte franchement, nous hésitons, nous envisageons les deux hypothèses en alternance. Dans le second, nous n'avons que des intuitions très faibles vis-à-vis de la décision à prendre : la situation est nouvelle, nous ne savons que faire. Dans ces deux cas, nous avons des chances de prendre une mauvaise décision – une décision qui puisse être perçue au moins comme un signe d'incompétence. Le fait d'être sûr de nous ne garantit bien entendu pas que nous allons prendre une bonne décision, mais nous nous tromperons nécessairement moins souvent dans ce cas que si nous sommes très incertains de la décision à prendre.

Prédictions sur le fonctionnement du raisonnement

A travers toutes ces situations, le raisonnement devrait conserver un fonctionnement relativement similaire (sinon on ne pourrait pas dire qu'il s'agit vraiment d'une seule capacité). Comme expliqué dans la partie précédente, cette activité revient à évaluer des arguments potentiels soutenant une conclusion donnée. Dans le cas d'une discussion, il est clair que la conclusion est le message que nous voulons transmettre. Dans le cas de la prise de décision, la conclusion prend une forme proche de 'la décision D est/était une bonne décision'. Il s'agit alors de trouver des arguments soutenant cette position. Dans les deux cas (discussion et prise de décision), il peut s'agir d'arguments soutenant directement la conclusion, d'argument attaquant les alternatives, ou d'arguments comparatifs montrant la supériorité de la conclusion face aux alternatives. Nous avons vu plus haut que le raisonnement avait tendance à faire du satisficing, à se contenter d'une solution assez bonne, avec des critères assez laxistes. Un des effets principal que va avoir le public est de modifier ces critères, à la fois qualitativement et quantitativement. Le fait que la décision soit plus ou moins publique, ainsi que l'importance qu'a le public à nos yeux aura des effets principalement quantitatifs : les critères seront plus élevés. La composition du public aura, elle, des effets qualitatifs : des arguments différents seront recherchés et retenus. Dans le cas de la discussion, les effets du public peuvent jouer à deux niveaux. On cherche avant tout des arguments qui soient efficaces dans leur objectif primaire de conviction. On peut également voir cette tentative de conviction comme une décision et il ne faut alors pas utiliser des arguments qui, bien qu'ils puissent être efficaces, risqueraient d'avoir de mauvaises conséquences pour notre réputation (que ce soit vis-à-vis d'un public externe au débat, ou même de notre interlocuteur). Par exemple un ami nous fait part de sa décision de se lancer dans le surf. Nous pensons qu'il s'agit d'une mauvaise idée, mais nous ne pensons pas que lui dire aura un effet, il nous faut des arguments. Le fait qu'il soit à peine capable de tenir l'équilibre sur un vélo nous vient immédiatement à l'esprit, et il s'agirait certainement d'un bon argument, mais nous peut-être ne nous risquerons-nous pas à le suggérer par peur de blesser cet ami maladroit.

Lorsque nous cherchons des arguments pour une décision que nous n'avons pas encore prise, ou pour une affirmation que nous n'avons pas encore énoncée, il est

possible de ne pas se contenter de chercher des arguments pour une seule option. On peut examiner tour à tour les arguments potentiels pour différentes options qui sont pour nous assez proches, au moins au sens ou aucune n'est clairement favorisée par rapport à l'autre (dans le cas des intuitions fortes mais contradictoires, elles ne sont proches que dans ce sens et pas dans le sens d'avoir des résultats similaires).

Prédictions sur les effets du raisonnement

Sur la base des prédictions concernant le fonctionnement du raisonnement, on peut tenter d'anticiper certains de ses effets. Il y a aura nécessairement une grande variance en fonction des circonstances dans lesquelles il est utilisé, mais on peut essayer de dégager des grandes lignes. Une distinction essentielle est celle du caractère anticipatoire ou rétroactif du raisonnement : est-ce que nous sommes en train de chercher à justifier une affirmation déjà énoncée, une décision déjà prise, ou au contraire sommes-nous en train de raisonner avant d'affirmer ou avant d'agir ? Dans le cas des utilisations rétroactives du raisonnement, il peut avoir deux effets différents. Le plus courant, de loin, est de renforcer notre croyance qu'il s'agit d'une bonne affirmation ou d'une bonne décision. Il est trop tard pour changer d'avis, le raisonnement ne peut donc que nous aider à justifier ce qui est déjà fait, et si nous parvenons en effet à trouver des arguments, cela ne peut que renforcer notre confiance.

On peut aussi concevoir le cas dans lequel il nous est impossible de trouver de bons arguments justifiant notre décision, ou étayant notre affirmation. Dans ce cas, il est possible que la recherche infructueuse nous amène au contraire à reconnaître qu'il s'agissait en fait d'une mauvaise idée – même si cela est loin d'être automatique : il nous arrive à tous de conserver foi en une idée en dépit du fait que nous n'arrivons pas à la défendre de façon convaincante. Cela a beaucoup plus de chances d'arriver à l'issue d'une conversation au cours de laquelle plusieurs de nos arguments ont déjà été réfutés qu'immédiatement après une décision : il est rare qu'on soit tout à fait incapables de trouver ne serait-ce qu'un seul argument pour soutenir une conclusion ou une décision donnée. Or l'immense majorité des situations expérimentales ne fait justement pas appel à de telles situations de discussion (étant donné que même les entretiens sont rarement confrontationnels,

l'exception principale étant celle du raisonnement en groupe), et on a donc des raisons de penser que dans la majorité des situations expérimentales, lorsque les participants doivent raisonner après avoir pris une décision ou énoncé une conclusion, l'effet principal sera un renforcement de la confiance en cet acte initial.

Des effets similaires devraient être observés lorsque les gens savent qu'ils vont être questionnés sur leur décision quelle que soit leur confiance initiale. Nous venons de voir que le raisonnement devrait se déclencher naturellement lorsque nous ne sommes pas trop sûrs de nous, mais beaucoup plus difficilement dans le cas contraire. Cependant il arrive que nous sachions que nous allons devoir justifier nos décisions même si nous sommes parfaitement sûrs de nous. Ces situations dans lesquelles nous sommes 'accountable', dans lesquelles nous devons rendre des comptes, ont été beaucoup étudiées expérimentalement car elles sont assez fréquentes dans certains contextes institutionnels. Si nous avons de fortes intuitions sur la décision à prendre, mais que nous raisonnons néanmoins, alors l'effet principal sera un renforcement de notre confiance car nous serons fortement motivés pour trouver des arguments soutenant cette décision. On peut cependant envisager certains cas dans lesquels le raisonnement nous fera tout de même changer d'avis. Les facteurs importants sont alors : (i) l'accessibilité et la force d'arguments soutenant la position initiale, (ii) l'accessibilité des options alternatives, (iii) l'accessibilité et la force d'arguments pour et contre les options alternatives et (iv) l'attrait intuitif relatif de la position initiale et des positions alternatives.

Si nous pouvons trouver facilement des arguments favorisant notre position initiale, il est bien possible que nous nous en satisfaisions. Dans ce cas, le raisonnement risque fort de n'avoir pour effet que d'augmenter notre confiance dans la décision initiale. Ce qui constitue les options alternatives sera parfois assez clair. Il peut s'agir d'options que nous pourrions considérer nous-mêmes, ou d'options que nous pensons que les autres (les personnes à qui nous allons devoir rendre des comptes) pourraient envisager. Dans ce cas, on peut commencer à raisonner, mais étant donné que nous sommes fortement enclins à préférer notre propre choix, le raisonnement sera principalement orienté contre ces options alternatives. S'il est facile de trouver des arguments soit portant directement contre ces dernières, soit confirmant la supériorité de notre position, le raisonnement risque à nouveau de n'avoir pour effet que d'augmenter notre confiance. Cependant, si nous nous trouvons incapables de découvrir des arguments contre une option alternative, il est

possible que nous finissions par changer d'avis, en particulier s'il est non seulement difficile de trouver des arguments contre cette option, mais que des arguments la soutenant sont facilement accessibles (la force des arguments étant ici en partie évaluée par des critères dépendants du public auquel nous devrons rendre des comptes). Dans ce cas, on peut dire que le raisonnement a joué un *rôle causal* dans la décision. Dans son usage normal le raisonnement n'a pas réellement d'influence sur la position que nous défendons (ou alors légèrement, sur la formulation d'une conclusion par exemple, mais pas substantiellement) : son but n'est pas de nous aider à former une position nouvelle, mais est de trouver des arguments la soutenant. Dans le cas qui vient d'être présenté cependant, le raisonnement peut influer, indirectement, sur la décision qui sera prise.

Ce rôle causal, le raisonnement le jouera beaucoup plus facilement lorsqu'il est activé pour des raisons que nous pouvons qualifier d'internes : non parce que nous avons la certitude d'avoir à rendre des comptes, mais parce que nous avons des intuitions faibles ou contradictoires. Dans ce cas, le raisonnement pourra plus aisément jouer un rôle 'd'arbitre', favorisant une décision car elle peut plus facilement être justifiée. On retrouve ici les facteurs qui jouent un rôle lorsque nous devons rendre des comptes, avec deux différences dans la façon dont ils sont considérés. D'une part, l'examen des arguments, quels qu'ils soient, sera généralement moins exigeant car nous ne sommes pas sûrs de devoir justifier notre décision. Il s'agit plus ici de s'assurer que nous pourrons dire quelque chose, pas nécessairement que l'argument sera très bon – bien qu'il puisse arriver que pour certaines décisions à caractère très public, ou très importantes, la pression soit telle qu'elle nous pousse à être aussi exigeant que si nous étions sûrs de devoir rendre des compte. La deuxième différence, de taille, est l'attitude que nous allons avoir vis-à-vis des options alternatives. Lorsque nous raisonnons surtout de par des motivations externes (nécessité de rendre des comptes), le point de vue initial sur les options alternatives sera négatif : notre objectif premier est de trouver des arguments soutenant notre position initiale. Mais si les motivations sont internes (intuitions faibles ou contradictoires), l'objectif est plus de parvenir à un compromis entre une option qui serait favorisée, même légèrement, par nos intuitions, et une option qui serait facile à justifier. Il convient dès lors plutôt de chercher des arguments soutenant les différentes options d'une manière similaire : non pas chercher des arguments contre, mais chercher des arguments pour chacune. Cela est vrai au moins

dans un premier temps ; une fois que la balance commence de pencher d'un côté, il est possible de revenir à une stratégie visant à montrer que l'option qui est alors favorisée est supérieure aux autres.

Lorsque le raisonnement est ainsi utilisé, son effet principal devrait-être de nous amener à prendre des décisions que l'on peut facilement justifier – pas forcément de meilleures décisions en elles-mêmes. Au contraire, l'effet net du raisonnement devrait même être de nous faire prendre de moins bonnes décisions : si on admet que la fonction de nos mécanismes intuitifs est justement de nous faire prendre de bonnes décisions, l'influence du raisonnement devrait avoir tendance à être néfaste – son objectif n'est pas de nous faire prendre de bonnes décisions, mais de nous faire prendre des décisions justifiables. Il arrivera cependant que le raisonnement nous fasse en fait prendre de meilleures décisions de façon indirecte : on peut penser en effet qu'une décision plus facilement justifiable a parfois plus de chances d'être bonne, et ce pour deux raisons relativement indépendantes. La première est de nature 'culturelle' : nous nous reposons souvent sur des justifications qui ne sont valables que parce qu'elles sont culturellement acceptées. Or il semble raisonnable de penser qu'un si un argument est culturellement accepté il a des chances d'être, en effet, valide : si la culture est filtrée par nos mécanismes cognitifs, et que ceux-ci sont globalement bien adaptés, il serait surprenant qu'ils laissent se répandre une majorité de croyances fausses. Il y a cependant tellement d'exceptions que cette généralité n'est que d'une utilité limitée lorsqu'il s'agit de l'examen d'un cas particulier. Ce qui détermine alors l'efficacité du raisonnement est à la fois la justesse de la justification choisie, et le fait qu'elle soit bien appropriée à la situation donnée.

Le raisonnement peut cependant aider à prendre de meilleures décisions même en l'absence d'éléments culturels – on pensera par exemple aux tâches classiques de raisonnement, telle que la tâche de sélection de Wason dans sa forme abstraite originale, dans lesquelles il est indéniable que le raisonnement est ce qui permet à certains participants de trouver la bonne réponse. Dans ces cas, l'effet du raisonnement est de déceler une erreur dans nos mécanismes intuitifs et de proposer une réponse supérieure, mais il est important de souligner que même alors, il ne s'agit que d'un effet de la recherche de réponses justifiables. Pour que cela soit possible, il faut que plusieurs conditions soient réunies. Il est tout d'abord indispensable – et ce n'est pas une condition négligeable – que nos mécanismes

intuitifs nous induisent en erreur : s'ils nous proposent d'emblée la bonne solution, le raisonnement ne pourra faire qu'empirer les choses. Il faut également que les options alternatives soient assez facilement accessibles. Si ce n'est pas le cas, le raisonnement ne fera que chercher des arguments pour la première et seule réponse envisagée. Et il faut finalement que les réponses alternatives soient plus facilement justifiables que la réponse initiale. Si ces conditions sont réunies, il est alors possible que la réponse finalement choisie, celle qu'on peut justifier, soit supérieure à la réponse initiale.

4 La vérification de cohérence

Dans le scénario évolutionniste qui a été esquissé plus haut, une étape précédant le raisonnement est celle de la vérification de cohérence. Il s'agit de comparer les informations qui nous sont communiquées aux informations que nous possédons déjà, qu'il s'agisse de croyances ou de plans (voir chapitre 1). Ce mécanisme est très prudent : il rejette trop d'informations, mais permet ainsi à l'individu d'être presque sûr de ne pas se faire manipuler. Cependant, chez l'humain de nombreux mécanismes ont évolué depuis qui permettent d'accepter plus d'informations, des mécanismes de confiance et de raisonnement en particulier. Grâce à eux, nous pouvons accepter des informations incohérentes avec ce que nous pensions ou voulions initialement. Il peut donc être difficile de montrer que ce mécanisme plus primitif de vérification de cohérence est toujours en activité. Une solution est d'essayer de trouver des cas dans lesquels les mécanismes plus sophistiqués ne sont pas utilisés : il sera alors peut-être possible d'observer le fonctionnement des mécanismes de vérification de cohérence.

Un bon moyen de tester l'existence de ce mécanisme est d'utiliser l'influence subliminale. Le principe général en est le suivant : les participants sont confrontés à des stimuli qui peuvent avoir une influence sur eux dont ils ne sont pas conscients. Dans ce cas, on peut penser que les participants ne pourront pas utiliser de mécanismes de raisonnement pour évaluer ce qu'on leur présente. Il n'est pas dit que les mécanismes de calibrage de la confiance ne puissent pas non plus être utilisés dans ces cas, mais on peut penser qu'ils le seront moins en tout cas. La théorie argumentative prédit que dans ce cas les mécanismes de vérification de cohérence seront actifs et que, étant donné qu'ils sont plus stricts que les mécanismes plus sophistiqués, les participants seront *moins* influencés par des stimuli subliminaux qu'ils ne le seraient pas des stimuli présentés consciemment (et assortis, si besoin, d'arguments).

Il n'y a pas, à ma connaissance, d'expérience comparant directement les effets de l'influence subliminale aux effets de signaux perçus consciemment (en partie car il serait difficile de trouver de bons contrôles : trouver l'équivalent conscient d'un stimulus subliminal, les matcher, etc.). On en est donc réduit à la prédiction suivante : les stimuli subliminaux devraient avoir une influence très

réduite sur notre comportement lorsqu'ils sont incohérents avec nos croyances et préférences.

4.1 Influence et persuasion subliminale

L'histoire de l'influence subliminale commence en 1957 lorsqu'un publicitaire fait scandale en prétendant avoir considérablement augmenté les ventes de pop-corn et de coca-cola en flashant ces mots de façon subliminale durant un film. Ces résultats seront repris dans le livre The Hidden Persuaders (Packard, 1957) qui prétendait montrer que nous étions en permanence influencés par de tels signaux subliminaux. Cependant, « l'expérience » s'est révélée avoir été une fumisterie montée pour relancer une entreprise de publicité (Weir, 1984). Par la suite, des expériences contrôlées n'ont trouvé aucun effet de tels stimuli subliminaux (Moore, 1982), ni des cassettes de développement personnel supposées nous aider durant notre sommeil (Greenwald, Spangenberg, Pratkanis, & Eskenazi, 1991). En 1992, Praktanis et Aronson passent en revue la littérature sur le sujet et concluent que l'influence des stimuli subliminaux n'a pas encore été fermement établie (Pratkanis & Aronson, 1992).

Depuis le début des années 1990, la situation a quelque peu changé. Plusieurs psychologues sociaux – John Bargh en particulier – ont démontré de nombreux effets de stimuli subliminaux sur le comportement. Dans une de ces expériences par exemple, certains participants devaient déchiffrer des anagrammes dont le résultat pouvait être des mots apparentés à la vieillesse. Il a été observé que ces participants marchaient moins vite, en sortant du laboratoire, que ceux qui avaient déchiffré des anagrammes de mots neutres (Bargh, Chen, & Burrows, 1996). Dans une optique plus proche de la persuasion, d'autres expériences ont montré que des stimuli flashés de façon subliminale pouvaient influencer le comportement. Par exemple, des participants à qui étaient flashés des mots liés à la soif buvaient plus de liquide à l'issue de l'expérience (Strahan, Spencer, & Zanna, 2002, voir également Berridge & Winkiehnan, 2003, pour un résultat similaire). Ces effets sont cependant modérés par les intentions et désirs préalables des participants. Dans le cas de la boisson par exemple, seuls les participants qui étaient déjà assoiffés furent influencés par les stimuli subliminaux : ces stimuli n'eurent aucun effet sur les participants qui

n'avaient pas soif. Voilà la façon dont Bargh explique la différence entre ces résultats et les échecs qui avaient été obtenus jusque là : « The main reason for the recent success is that researchers are taking the consumer's (experimental participant's) current goals and needs into account. » (Bargh, 2002, pp.282-3) Ou plus précisément :

> We suspect that previous attempts at subliminal persuasion have not harnessed the power of subliminal priming techniques in seeking to persuade people. In particular, we propose that subliminal priming can be used to prime goal-relevant cognitions, and that this priming when combined with a motive to pursue the goal will make persuasive appeals targeting this goal particularly effective. (Strahan et al., 2002, p.557).

En effet, ces expériences récentes d'influence subliminale ne déclenchent pas une nouvelle intention, elles ne font que modifier (assez légèrement généralement) la façon dont elle est menée à bien. Dans le cas de la soif, elles ne provoquent pas le désir de boire, elles modifient légèrement l'intensité de ce désir. Dans le cas de la vitesse de marche, ce ne sont pas les influences subliminales qui font que les participants décident de marcher pour quitter le laboratoire. Là encore, elles ne font que modifier la façon dont cette intention est menée à bien. Ces remarques s'appliquent aux autres expériences de ce domaine (voir Dijksterhuis & Bargh, 2001, pour revue). Cela ne signifie pas que ces résultats ne sont pas intéressants, ou qu'il est impossible que des publicitaires (par exemple) tirent profit de tels effets pour nous influencer à notre insu. Mais la conclusion générale doit en être que ces effets sont très limités, et qu'il ne semble pas être possible d'induire un comportement nouveau de cette façon, un comportement qui ne soit pas cohérent avec les intentions préalables de l'individu (voir également Dijksterhuis, Aarts, & Smith, 2002).

4.2 Repérer les énoncés incohérents avec nos croyances : le point de vue développemental

Nous venons de voir que dans le cas des intentions, les prédictions de la théorie argumentative sont vérifiées. Qu'en est-il des croyances ? Peu d'expériences

ont directement abordé ce sujet chez les adultes, sûrement car le résultat semble être relativement trivial : nous sommes tous capables de reconnaître qu'un énoncé est (directement) contraire à nos croyances. On peut cependant se poser la question de la détection de telles incohérences par de jeunes enfants. Il n'est pas évident que de jeunes enfants prêtent attention à de telles contradiction : après tout, une grande partie des informations qu'ils doivent acquérir est contraire à leurs intuitions – le fait que la Terre est ronde pour prendre un exemple bien étudié. Ces expériences sont d'autant plus intéressantes pour nous qu'il est probable que chez les enfants les mécanismes de vigilance épistémique plus sophistiqués (comme le raisonnement) soient moins développés. Cela nous donne donc une fenêtre vers une version plus 'pure' du mécanisme de vérification de cohérence (bien que les mécanismes basés sur la confiance soient actifs très tôt ; Clément, Koenig, & Harris, 2004).

Les premières expériences s'étant penchées sur la façon dont les enfants traitent les incohérences avaient recours au paradigme suivant. Une histoire est lue aux enfants ou, s'ils sont assez âgés, ils la lisent eux-mêmes. Parmi les énoncés constituant l'histoire, un est incohérent soit avec un énoncé précédent (incohérence textuelle), soit avec les croyances préalables de l'enfant (contre-vérité). Le problème, pour ce qui nous concerne, des incohérences textuelles est que pour que le mécanisme de vérification de cohérence s'y applique, il faut que la première information ait été entièrement intégrée à nos croyances. La fonction de ce mécanisme primitif de vérification de cohérence n'est en effet pas de vérifier que les autres sont cohérents (dans ce qu'ils font ou disent), mais de rejeter des informations incohérentes avec nos propres croyances ou intentions. Or le statut des croyances communiquées dans le cadre d'une histoire n'est pas clair : il est fort possible que les enfants ne les intègrent pas directement à leur stock de croyance. Dans ce cas, le mécanisme de vérification de cohérence de base ne détecterait pas des incohérences internes à la narration (il est cependant possible que des enfants assez âgés puissent détecter de telles incohérences par le biais de mécanismes plus sophistiqués).

Qu'observons-nous dans le cas des contre-vérités (des énoncés contraires aux croyances des enfants) ? Une première expérience de Markman et Gorin semble donner des résultats plutôt négatifs (Markman & Gorin, 1981). Dans cette expérience, des enfants de 8 et 10 ans devaient lire des histoires dont certaines contenaient des contre-vérités. Sans instructions spécifiques concernant ces contre-vérités, les enfants avaient de mauvaises performances, ne détectant en moyenne

qu'une contre-vérité sur quatre. Par opposition, une expérience de Vosniadou et collègues (Vosniadou, Pearson, & Rogers, 1988) a montré que des enfants de 6 ans étaient capables de détecter des problèmes similaires (détection de plus de la moitié des contre-vérités). Pour expliquer cette différence il faut se tourner vers la façon dont les enfants étaient interrogés. Dans l'expérience de Markman et Gorin, ils devaient se prononcer sur la facilité de compréhension de l'histoire. Il s'agit d'une question trop vague. Il est tout à fait possible que les enfants aient simplement rejeté l'information fausse. Etant donné la façon dont les histoires étaient construites, éliminer l'information fausse rendait le reste de l'histoire très facile à comprendre. Vosniadou, elle, demandait directement aux enfants ce qui n'allait pas dans l'histoire. Cela peut sembler être une aide trop importante, mais si les enfants ne faisaient pas la différence entre les énoncés faux et les autres, ils devraient choisir un énoncé au hasard. Etant donné que chaque histoire contenait plus d'une dizaine d'éléments, le fait qu'ils soient capables de repérer les contre-vérités dans plus de la moitié des cas indique en fait une très bonne performance (pour les enfants de 6 ans, ceux de 9 ans étant quasiment au plafond).

Ces tâches sont malgré tout assez complexes, et plus récemment des chercheurs ont montré que des enfants beaucoup plus jeunes étaient capables de détecter les incohérences – et de préférer leurs propres croyances. Ainsi, dans une expérience de Clément, Koenig et Harris, les enfants étaient confrontés à la situation suivante (Clément et al., 2004). Un expérimentateur manipule deux poupées. La première est décrite comme étant fiable, et la seconde comme ne l'étant pas, puis ces qualités sont démontrées aux enfants : lorsqu'un objet est présenté, la première le nomme ou le décrit correctement, alors que la seconde se trompe à chaque fois. A cette première phase succèdent plusieurs tests de la compréhension des enfants, dont le plus intéressant ici est le dernier. Dans ce test, un pompon coloré est posé sur une boîte pour que l'enfant puisse bien le voir, avant d'être placé à l'intérieur de la boîte, maintenant hors de vue de l'enfant. Alors les deux poupées donnent de mauvaises couleurs pour décrire le pompon. L'expérimentateur demande alors à l'enfant la couleur du pompon. Dès 3 ans, les enfants répondent majoritairement que la couleur du pompon est celle qu'ils ont vu, et non celle donnée par une des poupées – même par la poupée qui jusque là avait toujours été parfaitement fiable. Dans ce cas donc on voit que les enfants même très jeunes favorisent leurs propres croyances (en tout

cas lorsqu'elles résultent de la perception) par rapport à celles qui leur sont communiquées.

4.3 Le biais égocentrique

Dans le cas des croyances, le mécanisme de vérification de cohérence ne doit pas uniquement nous faire réaliser qu'il y a un conflit, mais il devrait également favoriser nos propres croyances au dépend des croyances communiquées avec lesquelles elles entrent en conflit (ce qu'on peut appeler un *biais égocentrique*). Le fait que les adultes disposent de nombreux autres mécanismes pour évaluer les informations communiquées rend l'observation d'un tel phénomène ardue. On peut peut-être l'observer lorsque les participants sont peu motivés pour utiliser des mécanismes plus sophistiqués. Dans les termes d'une des théories de la persuasion, cela correspond aux cas dans lesquels « la probabilité d'élaboration est faible », et on observe alors en effet que : « la position du message pourrait servir d'indice simple. Les messages agréables seraient acceptés, mais les messages désagréables seraient rejetés en n'étant que peu analysés » (Petty & Wegener, 1998). Un tel effet a été observé par Petty et Cacioppo (1979) : lorsque l'intérêt des participants pour le message est limité, ils se fient principalement à sa direction (pro- ou contre-attitudinalle[29]) pour l'évaluer. Or c'est précisément cette direction qui est évaluée par les mécanismes de vérification de cohérence (voir Mullainathan & Shleifer, 2005, pour une application originale de ce phénomène au monde de la publicité).

Une autre façon d'évaluer le biais égocentrique est d'essayer de contrôler les autres facteurs pouvant jouer un rôle dans l'évaluation : si la personne qui nous transmet une information est aussi compétente que nous et que nous pouvons lui faire confiance, les mécanismes plus sophistiqués d'évaluation devraient rendre un verdict neutre. Si un biais persiste, il vient alors peut-être du mécanisme plus primitif de vérification de la cohérence. Certaines expériences ont étudié précisément ce cas. Les participants doivent donner une réponse initiale à une question donnée, par exemple sur la date d'un événement historique, puis on leur donne un conseil (en leur disant que la source est aussi compétente qu'eux), et ils peuvent alors donner une nouvelle

[29] C'est-à-dire allant soit dans le sens (pro) soit à l'encontre (contre) des attitudes préalables des participants.

réponse qui prenne en compte ce conseil. On observe encore un fort biais égocentrique ('egocentric advice discounting'). Si on place la réponse initiale du participant à 0 et celle du conseil à 100, la réponse moyenne après prise en compte du conseil est proche de 30 (Harvey & Fischer, 1997; Harvey, Harries, & Fischer, 2000; Mercier, Van der Henst, Yama, Kawasaki, & Adachi, submitted; Van der Henst, Mercier, Yama, Kawasaki, & Adachi, 2007; Yaniv & Kleinberger, 2000).

Cet effet pourrait être le résultat d'autres mécanismes et non d'un biais évolué. Il a ainsi été proposé qu'il ne s'agisse que d'un simple ancrage (Tversky & Kahneman, 1974) : étant donné que le récepteur forme habituellement son opinion avant qu'il ne reçoive le conseil, il pourrait s'en servir comme d'une ancre – c'est-à-dire d'un point de référence initial. Mais l'effet s'observe également lorsque le conseil est donné avant que le récepteur ne puisse former une opinion : l'ancrage ne peut donc totalement rendre compte de ce biais. Ilian Yaniv a quant à lui proposé que le poids accordé à sa propre opinion résulte d'un meilleur accès aux raisons nous ayant poussé à la former : nous connaissons quelques une des raisons ayant généré notre opinion alors que dans les contextes expérimentaux utilisés nous n'avons pas accès aux raisons ayant poussé les autres à former leur opinion (Yaniv, 2004; Yaniv & Kleinberger, 2000). Cette explication ne peut pas non plus être valable dans tous les cas de biais égocentrique : dans certains cas les récepteurs n'ont aucune raison qui pourrait justifier leur opinion, et le biais s'observe encore (Krueger, 2003). Krueger conclut qu'il s'agit donc bien d'un biais égocentrique qui nous fait préférer notre opinion simplement parce que c'est la nôtre.

Un autre cas dans lequel le biais égocentrique pourrait avoir des effets est celui des inférences. Lorsque quelqu'un nous dit quelque chose, que nous l'acceptons et que nous faisons des inférences en nous basant sur ces informations, ces nouvelles conclusions sont considérées comme les nôtres. Il doit donc être possible de comparer ce cas à celui où ces mêmes conclusions nous sont communiquées : nous devrions être biaisés vers les conclusions que nous avons tirées nous même. Une façon de tester cette prédiction est d'évaluer l'impact d'arguments dont la conclusion est explicite ou implicite. Il s'agit d'arguments en tous points similaires, à l'exception de la conclusion : alors que dans le premier cas la conclusion ('Donc la peine de mort devrait être interdite', par exemple) est mentionnée, dans le second elle ne l'est pas, bien que tous les arguments pointent dans sa direction. Les recherches dans ce domaine (passées en revue dans Petty & Wegener, 1998) ont tout d'abord

donné des résultats contradictoires, certaines montrant plus de pouvoir persuasif lorsque la conclusion était explicite et d'autres lorsqu'elle était laissée implicite. Des études plus récentes donnent l'avantage aux conclusions implicites, dans la mesure où les participants sont motivés et capables de les tirer par eux-mêmes (voir Stayman & Kardes, 1992). On peut donc dire que moins l'émetteur joue de rôle dans la formation de la conclusion par le récepteur, plus celui-ci l'acceptera facilement. Utiliser un maximum d'éléments implicites est une tactique un peu risquée, car dans certains cas le récepteur ne dérivera pas la conclusion souhaitée, mais elle peut être très efficace et certains auteurs l'utilisent à leur profit : « persuader les gens ne m'intéresse pas. Ce que j'aime c'est les aider à se persuader tout seul » (Chomsky & Barsamian, 2002, p.16).

Tous ces résultats convergent pour confirmer l'existence d'un biais égocentrique : toutes choses égales par ailleurs, nous aurons tendance à préférer les croyances auxquelles nous sommes parvenues par nous-mêmes à celles qui nous ont été communiquées.

4.4 Objections et réponses

L'aspect quasi automatique de la vérification de cohérence a été remis en question, et ce tout d'abord parmi les chercheurs étudiant la compréhension de texte, dont certains pensent que nous ne sommes pas toujours capables de résoudre les contradictions présentes dans un texte (Otero & Kintsch, 1992). Mais on peut objecter que ces études ne portent pas directement sur le mécanisme de vérification de cohérence qui nous intéresse ici : les textes utilisés ne contiennent pas de contre-vérité, uniquement des contradictions textuelles. Par exemple, un texte utilisé mentionne au début que la supraconductivité a seulement été atteinte à de très basses températures, et à la fin qu'elle a été atteinte en élevant les matériaux à de très hautes températures. Etant donné le caractère très abscons du sujet, il n'est guère surprenant que les participants n'aient prêté qu'une attention assez superficielle au texte, et n'aient pas incorporé son contenu directement à leurs croyances (ce qui est démontré par le fait que 10% des participants ne se souvenaient d'aucun élément du texte lors d'un rappel qui suivait presque immédiatement sa lecture). De plus, lorsqu'on s'intéresse aux réponses, la contradiction n'apparaît jamais telle quelle parmi ce que

les participants peuvent restituer du texte. Certains ne mentionnent simplement aucun des deux énoncés contradictoires, d'autres n'en mentionnent qu'un, et enfin d'autres ont résolu la contradiction d'une façon 'créative' (en disant par exemple que les techniques ont changé). Il est particulièrement intéressant de noter que dans ce cas un bon nombre de participants a eu recours à la stratégie la plus simple possible, celle résultant de l'utilisation simple de la vérification de cohérence : ils n'ont pas mémorisé un des deux énoncés contradictoires – ce qui est à attendre de la part de participants peu intéressés par le sujet. La conclusion des auteurs n'est donc pas du tout en contradiction avec la thèse défendue ici. Si on ne peut pas dire qu'il s'agisse d'une résolution 'intelligente' de la contradiction, la stratégie alternative utilisée par les participants est précisément celle qui est prédite par l'utilisation des mécanismes de vérification de cohérence : « the model assumes that suppression of contradictions is an inherent, normal part of comprehension. Here we wish to explore the conditions under which this normal suppression process can go awry and suppress actual text propositions. » (Otero & Kintsch, 1992, p.229).

D'autres recherches aux conclusions apparemment incompatibles avec la théorie proposée ici ont été conduites par Gilbert et ses collègues au début des années 90. Ils défendent une théorie 'spinoziste'[30] des processus de compréhension : pour eux, comprendre un énoncé implique son acceptation automatique, et la remise en cause n'est qu'un processus ultérieur facultatif. Ils s'opposent à une théorie 'cartésienne' selon laquelle la compréhension précéderait et serait différente de l'acceptation. D'un point de vue évolutionniste, des mécanismes fonctionnant de la façon décrite par la théorie spinoziste exposent à certains risques : si n'importe quel événement vient perturber la phase d'évaluation, nous nous retrouvons avec une croyance non examinée. A l'inverse, dans la théorie cartésienne, une perturbation aurait plutôt tendance à entraîner l'oubli de la croyance : elle n'a pas été acceptée, et n'a pas encore pénétré notre stock de croyance. Cela paraît beaucoup plus prudent : on pourrait penser que si la stratégie spinoziste était effectivement employée, des individus auraient trouvé un moyen d'en profiter. Il ne s'agirait dès lors plus d'une stratégie viable, et elle aurait dû être abandonnée. Quoiqu'il en soit, il convient

[30] Il semble qu'on puisse trouver des antécédents chez Aristote, pour qui « It is the mark of an educated mind to be able to entertain a thought without accepting it », Ce qui implique que les esprits non éduqués ne parviennent pas à faire cette distinction.

d'examiner les expériences que Gilbert et ses collègues avancent pour défendre leur thèse.

La première de ces expériences servira d'illustration (Gilbert, Krull, & Malone, 1990). On fait croire aux participants qu'ils vont prendre part à une expérience sur l'apprentissage du langage. Des énoncés, censé servir à apprendre des mots de Hopi, sont présentés aux participants, tels que « A monishna is a star ». Certains de ces énoncés sont accompagnés d'une mention indiquant s'ils sont en fait vrais ou faux. Pour certains énoncés, le traitement est perturbé par une tâche distractive (appuyer sur un bouton rapidement après avoir entendu un son). Les participants remplissent ensuite une tâche de mémoire : on leur demande par exemple « Is a monishna a star? ». Les prédictions de Gilbert et ses collègues sont les suivantes. Si l'acceptation est une phase supplémentaire suivant la compréhension (théorie cartésienne), les énoncés dont le traitement a été perturbé devraient davantage être traités comme s'ils étaient faux (car la phase d'acceptation a été perturbée). Et l'inverse devrait être vrai si la phase supplémentaire permet au contraire de les rejeter (théorie spinoziste) : si cette phase est perturbée, alors plus d'énoncés devraient être acceptés et traités comme vrais. Le résultat important est le suivant : lorsque le traitement des énoncés a été perturbé, les participants avaient tendance à les juger comme vrais, confirmant donc la théorie spinoziste défendue par Gilbert.

Les autres expériences décrites dans cet article, ainsi que dans un article suivant (Gilbert, Tafarodi, & Malone, 1993) sont cohérentes avec la théorie spinoziste. Cependant, des expériences plus récentes ont fourni des résultats incompatibles avec cette dernière. La différence introduite par Hasson et collègues est la suivante : les énoncés dont ils se sont servis étaient informatifs même lorsqu'ils étaient faux (Hasson, Simmons, & Todorov, 2005). Dans le cas de l'expérience utilisant les mots Hopi de Gilbert et al., savoir qu'un des énoncés utilisés est faux est extrêmement peu informatif (savoir qu'étoile ne se dit pas monishna en Hopi est inutile). A l'inverse, certains énoncés utilisés par Hasson et al. sont informatifs quand ils sont faux (par exemple 'Jean a une télévision' : le fait de ne pas avoir de télévision est plus informatif que le fait d'en avoir une). Lorsque de tels énoncés sont utilisés (dans une expérience par ailleurs similaire à celle de Gilbert), l'effet disparaît complètement, et les énoncés dont le traitement a été perturbé sont aussi bien mémorisés qu'ils soient vrais ou faux. Il faut noter tout de même que cela ne fournit

pas non plus de soutien pour une théorie cartésienne forte, qui aurait prédit que dans ces conditions les énoncés vrais seraient moins bien mémorisés (ce qui n'était pas le cas). Dans leur deuxième expérience, Hasson et al. ont eu recours à un autre type de tâche. Les énoncés étaient suivis d'une tâche de décision lexicale dont certains mots étaient liés à l'énoncé, ou à sa négation. Dans certains cas, les participants apprenaient si l'énoncé était vrai ou faux avant la tâche de décision lexicale, alors que dans d'autres cas ils ne l'apprenaient qu'après. Dans ce dernier cas, la théorie spinoziste prédit que l'énoncé sera d'abord compris et perçu comme vrai, et donc que la réponse à la tâche de décision lexicale sera la même que lorsque le participant sait déjà que la phrase est vraie. Or c'est le résultat inverse qui a été obtenu : lorsque les participants ne savaient pas encore si la phrase était vraie ou fausse, les performances étaient similaires au cas où les participants savaient qu'elle était fausse, ce qui indique qu'ils la considéraient alors comme étant plutôt fausse. Ces résultats limitent donc les conclusions de Gilbert : elles ne sont valides que lorsque l'énoncé est très peu informatif lorsqu'il est faux.

Une autre critique potentielle est soulevée par Petty et Wegener : « However, if a source is known to be a liar, and this information is retrieved, statements from this source would not be assumed to be true even if no specific issue-relevant information to the contrary was available. » (Petty & Wegener, 1998). Malheureusement, il n'y a pas à ma connaissance d'expériences montrant que cet argument – qui est intuitivement très plausible – est en effet valide. Mais il n'y a pas non plus de preuve contraire : dans les expériences de Gilbert, le 'locuteur' n'est pas une source dont nous pourrions naturellement nous méfier – comme quelqu'un qui nous aurait menti de manière répétée dans le passé.

Plus important pour ce qui concerne le mécanisme de vérification de cohérence, une autre limitation des expériences de Gilbert est la suivante : « We used these "nonsense propositions" (rather than real propositions such as potatoes are grown in Idaho) simply to ensure that subjects would use the signal word, and not their prior knowledge, to evaluate the truthfulness of the proposition. » (Gilbert et al., 1990, p.603). Donc les résultats de Gilbert et al. ne concernent pas les cas dans lesquels l'information communiquée pourrait entrer en conflit avec ce que les participants croyaient précédemment.

On peut tirer de tout cela la conclusion suivante. Lorsqu'on est confronté à un énoncé E qui n'est pas incohérent avec nos croyances (car il est totalement nouveau

par exemple), qui vient d'une source dont on n'a pas de raison de se méfier, et qui ne serait pas pertinent s'il était faux, alors il faut fournir un effort pour le maintenir découplé (sous une forme 'il est faux que E', ou 'il n'est pas sûr que E'). Si cet effort est perturbé, la suite la plus probable est l'oubli à la fois de l'énoncé et de sa valeur de vérité. Si, plus tard, on nous rappelle cet énoncé, il réactive des traces en mémoire et nous nous souvenons l'avoir vu (ou entendu). Parmi les gens qui se souviennent de l'énoncé, certains parviendront peut-être à se souvenir qu'il était faux et pourront répondre correctement. Mais la plupart des gens ne s'en souviendront pas, et répondront qu'il était vrai – d'une certaine manière, leur système cognitif fait l'hypothèse que si on ne se souvenait pas de l'information comme étant fausse, c'est qu'elle devait être vraie. L'effet est donc dû à la très faible pertinence du fait que l'information soit fausse au moment ou elle est encodée.

On peut raisonnablement penser que ces conditions étaient très rarement réunies dans l'environnement dans lequel nous avons évolué : lorsque quelqu'un que nous connaissons nous dit quelque chose qui s'avère être faux, il s'agit là presque toujours d'une information intéressante car elle nous permet de calibrer la confiance que nous accordons à cette personne. Cette 'faiblesse' n'en était donc pas vraiment une, mais plutôt un effet sans conséquence dû au fonctionnement adaptatif de la mémoire. L'environnement a cependant beaucoup changé depuis cette époque, et il est possible que les conditions qui rendent cet effet néfaste soient plus souvent réunies dans notre environnement moderne. Par exemple, une grande partie des informations nous parviennent de sources anonymes, ou dont l'évaluation n'est pas naturellement pertinente car il ne s'agit pas de personnes avec lesquelles nous pourrions être amenés à interagir (la publicité étant un bon exemple).

5 Raisonnement et argumentation

Every one is practicing oratory on others thro the whole of his life.
Adam Smith, Lectures on Jurisprudence, p.iv

Pour la théorie argumentative, la fonction première du raisonnement est de produire et d'évaluer des arguments suite à un désaccord, ou en anticipation d'un désaccord. Cela signifie que le raisonnement devrait être le plus efficace dans ce type de situation, lorsque son activation est la plus naturelle : quand deux personnes (ou plus) tentent de résoudre un désaccord par le biais d'arguments. Il s'ensuit que le contexte dans lequel le raisonnement devrait fonctionner le mieux est le raisonnement en groupe. Nous verrons donc, dans la première partie de ce chapitre, des résultats qui montrent qu'en effet, dans ces situations de raisonnement en groupe, les gens sont capables à la fois de produire de bons arguments pour défendre leurs positions, mais également qu'ils sont sélectivement sensibles aux meilleurs arguments – le tout ayant pour effet de faire converger le groupe vers la bonne solution. Le reste du chapitre sera consacré aux études portant sur les différents aspects des capacités nécessaires à une argumentation efficace, dans des contextes généralement un peu moins naturels que celui du raisonnement en groupe, mais qui compensent ce défaut par une plus grande rigueur dans le contrôle des stimuli.

5.1 Le raisonnement dans son contexte le plus naturel : en groupe

5.1.1 Le raisonnement en groupe est efficace

Dans une tâche classique de psychologie du raisonnement, les participants sont confrontés à des énoncés et ils doivent essayer de former une conclusion logiquement valide sur la base de ces énoncés ou évaluer une conclusion proposée. Bien que le terme 'argument' soit employé en logique pour décrire les ensembles de propositions et de conclusions que les participants doivent produire ou évaluer, ces arguments n'ont rien d'argumentatif au sens courant du terme : on ne demande pas

aux participants s'ils acceptent ou non la conclusion – ce qui serait généralement absurde étant donné que le contenu est artificiel dans la plupart des cas – mais si l'argument est logiquement valide. Autrement dit, personne n'essaie d'argumenter avec les participants, ou de les persuader de quoi que ce soit. Dans ces circonstances, la théorie argumentative prédit que le raisonnement ne devrait qu'être difficilement activé et donc que les performances devraient être basses. C'est en effet ce qu'observe Evans dans sa revue de la littérature sur le raisonnement déductif : « it must be said that logical performance in abstract reasoning tasks is generally quite poor » (Evans, 2002, p.981). Durant ces tâches, les participants luttent avec des syllogismes (voir section 6.3 pour une introduction) ou des énoncés conditionnels (du type 'si p, alors q'). Par exemple, ils ne sont en moyenne que 60 % à donner la réponse logiquement correcte dans le cas du *modus tollens* (si p alors q, non q donc non p) (Evans, Newstead, & Byrne, 1993). De même, dans une autre tâche requérant la compréhension de la table de vérité du conditionnel – la fameuse tâche de sélection de Wason (Wason, 1966, voir section 3.1) – ils ne sont souvent que 10 % à trouver la bonne réponse (Evans, 1989; Evans et al., 1993; Griggs & Cox, 1983).

Une façon simple d'inscrire ces tâches dans un contexte argumentatif est de créer des groupes de participants et de leur demander de résoudre ces mêmes problèmes : pour peu que les participants ne soient pas tous d'accord dès le début, ils devraient alors débattre pour se mettre d'accord sur la meilleure solution. La littérature sur la prise de décision en groupe est immense (voir Kerr, Maccoun, & Kramer, 1996; Kerr & Tindale, 2004 pour des revues récentes), mais ici nous nous intéresserons surtout aux problèmes qui peuvent être comparés facilement avec ceux qui sont généralement utilisés en psychologie du raisonnement. Il faut donc se cantonner aux problèmes dit 'intellectifs', c'est-à-dire ceux pour lesquels « il existe une réponse dont on peut démontrer la véracité au sein d'un système conceptuel verbal ou mathématique » (Laughlin & Ellis, 1986, p.177), ou à ceux qui sont proches de l'extrémité intellective du continuum qui les oppose aux problèmes de jugement purs (problèmes dont la solution est entièrement subjective). Dans ce cas, le schéma (au sens de Davis, 1973) qui rend le mieux compte des données est celui de « la vérité gagne » : si un des participants a trouvé la solution et a la bonne justification, alors cette réponse l'emportera. Un tel phénomène a été observé dans des tâches mathématiques (Laughlin & Ellis, 1986; Stasson, Kameda, Parks, Zimmerman, & Davis, 1991), des problèmes de type 'Eureka' (Laughlin, Kerr,

Davis, Halff, & Marciniak, 1975), et des problèmes du jeu 'Mastermind' (B. L. Bonner, Baumann, & Dalal, 2002). Lorsqu'il s'agit de problèmes d'induction – dont la réponse correcte ne peut pas être parfaitement prouvée, mais peut tout de même être raisonnablement soutenue – le schéma le plus courant est celui de « la vérité soutenue gagne » dans lequel il faut qu'au moins deux personnes aient la bonne réponse, et la défendent, pour qu'elle l'emporte (Laughlin, Bonner, & Miner, 2002; Laughlin, Hatch, Silver, & Boh, 2006; Laughlin, VanderStoep, & Hollingshead, 1991; Laughlin, Zander, Knievel, & Tan, 2003). On observe donc que dans ces tâches les performances sont nettement supérieures : les groupes ont tendance à être au niveau de leur meilleur membre, et leur performance est donc nettement supérieure à la moyenne d'individus isolés (voir les références cités ci-dessus et Hill, 1982, pour une revue des travaux antérieurs).

Ces résultats fournissent de bonnes raisons de penser que les tâches typiquement utilisées par les psychologues du raisonnement – qui sont soit purement intellectives, soit très proches de cette extrémité du continuum – devraient elles aussi être résolues beaucoup plus facilement en groupe. Malheureusement, peu de tâches ont ainsi été testées en groupe mais l'exception principale répond parfaitement aux attentes[31]. Dans plusieurs expériences les participants ont eu à résoudre la tâche de sélection de Wason en groupe. Dans la première de ces expériences, Moshman et Geil (Moshman & Geil, 1998) ont procédé à deux comparaisons : des participants faisant face à la tâche de sélection de Wason seuls et des participants la résolvant en groupe, et des participants la résolvant d'abord seuls puis en groupe. Les participants s'attelant seuls à cette tâche avaient en moyenne 14 % de bonnes réponses. Par contre, 70 % des groupes ayant affronté directement la tâche et 80 % de ceux pour qui elle succédait à la passation individuelle sont parvenus à la bonne réponse. Une forte augmentation des performances est également notée par Maciejovsky et Budescu (2007) dans leur première expérience : 9 % des participants trouvent la bonne réponse seuls, mais 50 % des groupes y parviennent[32].

[31] Par ailleurs, de nombreux problèmes de raisonnement ne seraient pas idéaux pour des contextes de groupe car tous les participants tendent à faire la même erreur, auquel cas on n'attend pas de débat et donc pas d'amélioration.

[32] La différence de performance des groupes est moins importante, mais ceci est simplement dû au fait que dans cette seconde expérience moins de participants réussissaient la tâche par eux-mêmes, et donc plus de groupes ne comprenaient que des participants en accord sur la mauvaise réponse.

Réplication de Moshman et Geil 1998

Afin d'étudier de plus près les processus à l'œuvre lorsque les membres d'un groupe doivent résoudre un problème de raisonnement, j'ai mené à bien une réplication de l'expérience de Moshman et Geil (1998) sur la tâche de sélection de Wason. Les participants étaient des étudiants de première année de biologie, assistant à un TD de méthodologie (quatre classes en tout). Etant donné qu'il s'agissait d'une simple étude exploratoire, la méthode retenue fut celle permettant d'obtenir le plus de données : faire passer le test d'abord en individuel puis en groupe. Cette méthode permet d'étudier l'évolution individuelle des participants, et surtout permet de connaître la composition des opinions initiales au sein de chaque groupe. Le défaut est qu'on pourrait arguer qu'une éventuelle amélioration n'est due qu'à un effet d'entraînement. Cette explication est toutefois très peu plausible. Les participants ont beaucoup de temps pour résoudre la tâche en individuel, et il est très probable que leurs performances plafonnent, et qu'une nouvelle passation n'aurait aucun effet. C'est d'ailleurs ce qu'ont fait certains chercheurs en utilisant comme condition contrôle une répétition de la tâche en individuel. Dans ce cas, aucune amélioration n'était apportée par la répétition elle-même (voir les références citées plus haut sur la résolution de problèmes intellectifs en groupe). Par ailleurs, dans le cas de la tâche de sélection, Moshman et Geil avaient également utilisé une condition dans laquelle les participants ne résolvaient la tâche qu'une fois, soit en groupe soit en individuel : la différence entre ces deux groupes était similaire à celle observée lorsque les mêmes participants affrontaient le problème individuellement puis en groupe.

Les résultats de mon expérience furent similaires à ceux existant dans la littérature (Maciejovsky & Budescu, 2007; Moshman & Geil, 1998). Individuellement, 12 participants sur 58 trouvèrent la réponse correcte (21% de réponses correctes), alors que 7 des 14 groupes y parvinrent (50% de réponses correctes) (χ^2 (1)=4,98, p<0,5). Le fait que l'amélioration soit moins importante que chez Moshman et Geil est peut-être dû à la répartition des personnes ayant initialement les bonnes réponses (ou des éléments de bonne réponse) au sein des groupes : dans cette expérience, elles tendaient à se retrouver au sein des mêmes groupes, ce qui a réduit leur influence.

On peut néanmoins faire une analyse plus fine qui révèle l'efficacité du raisonnement en tant qu'outil pour évaluer des arguments. Les positions initiales des membres de chaque groupe étant connues, il est possible d'étudier l'évolution des réponses en fonction des arguments présents dans le groupe. On peut ainsi comparer le nombre de réponses qui ont été influencées par des arguments pour la réponse correcte au nombre de réponses ayant été influencées par des arguments pour la réponse incorrecte – et ce pour chacune des quatre cartes. Etant donné que la réponse qui concerne chaque carte (faut-il ou non la retourner) est correcte ou incorrecte, il suffit de compter le nombre de désaccords initiaux parmi les membres d'un groupe. Imaginons un groupe de deux personnes, la première à choisi de retourner p et non-q, alors que la seconde a elle choisi p et q. Dans ce cas, il y a deux conflits (sur q et sur non-q), sur un total de quatre possibles.

Il y eu au total 73 conflits de cette sorte. Quelle en fût l'issue ? Dans 9 cas (12%), c'est la personne ayant initialement donné la bonne réponse qui a changé d'avis, au profit de celle ayant donné la mauvaise. Mais dans la grande majorité des cas (63, soit 85%), c'est l'inverse qui c'est produit : la personne ayant initialement donné la mauvaise réponse a été convaincue par celle qui avait trouvé la bonne (une différence fortement significative : $\chi^2(1) = 40{,}5,\ p<0{,}00001$)[33]. De plus, pour avoir écouté certaines conversations, j'ai un sentiment clair que les quelques cas de personnes ayant modifié leur réponse vers une réponse incorrecte correspondent à des personnes qui avaient donné la bonne réponse pour de mauvaises raisons. Etant donné que chaque réponse est dichotomique, il est possible que certaines personnes aient choisi certaines bonnes réponses sans avoir une bonne raison. Nous verrons plus tard (section 5.2.1) que lorsque la tâche requiert que les participants trouvent la bonne raison pour avoir la bonne réponse, alors les personnes qui ont en effet la bonne réponse ne changent jamais d'avis.

[33] Le cas restant correspond à une personne n'ayant pas changé d'avis – malgré les instructions qui demandaient qu'un consensus soit atteint.

5.1.2 Explications alternatives

Avant de pouvoir interpréter ces résultats comme un soutien pour la théorie argumentative du raisonnement, il faut répondre à plusieurs objections. Peut-être que les améliorations ne viennent que d'un échange d'information, et qu'aucun raisonnement n'a lieu au sein du groupe ? Pour que cette objection soit valide, il faudrait que la personne qui a la bonne réponse ait simplement plus d'information que les autres membres du groupe, et qu'il lui suffise alors de les partager avec eux pour qu'ils acceptent la bonne réponse. Ceci ne décrit pas du tout la façon dont les sessions de groupe se déroulent. Plutôt qu'un simple échange d'information, on observe des débats et des arguments. Il suffit d'assister à une séance pour s'en convaincre, et cette impression est confortées par une analyse même superficielle des transcriptions disponibles dans la littérature (voir Moshman & Geil, 1998; Trognon, 1993, pour la tâche de sélection de Wason). Plus généralement, le rôle des conflits dans l'amélioration des performances en groupe a été démontré par de nombreux travaux (voir les nombreuses références cités dans Schulz-Hardt, Brodbeck, Mojzisch, Kerschreiter, & Frey, 2006). Enfin, les expériences dans lesquelles les participants doivent résoudre les tâches individuellement puis en groupe montrent qu'un simple échange d'information ne peut suffire. Dans le cas de la tâche de sélection de Wason, les membres du groupe arrivent généralement avec des solutions différentes, mais les solutions fausses ne sont pas seulement incomplètes : elles comprennent également des éléments incorrects. Pour parvenir à la solution correcte, les membres du groupe doivent non seulement accepter de retourner les cartes proposées par le ou les partisans de la réponse correcte, mais également rejeter celles qui ne sont pas pertinentes : ils doivent discriminer les informations qui leurs sont communiquées, accepter celles qui sont bien argumentées et rejeter celles qui ne le sont pas.

Une autre objection potentielle (soulevée par Oaksford, Chater, & Grainger, 1999, dans le cas de la tâche de sélection de Wason) est que les participants avec un fort QI – et donc ayant plus de chances de parvenir à la bonne solution dans ce type de problème[34] – exercent une influence démesurée sur le groupe. En d'autres termes,

[34] Il ne s'agit pas d'une prise de position forte sur la valeur du QI (loin de là), mais plutôt d'une quasi tautologie dans la mesure où les tâches mesurant le QI sont très proches des tâches étudiées ici.

les membres du groupe reconnaîtraient les 'experts' et suivraient leur avis. A nouveau, cette objection est très peu plausible. Comme nous venons de le voir, les participants ne font pas qu'accepter la solution avancée par un membre du groupe : ils débattent et argumentent, et leur performance est liée à ces débats. Plus fondamentalement encore, les participants ne se connaissent généralement pas avant d'interagir au sein du groupe : comment pourraient-ils alors reconnaître l'expert ? S'il y a bien reconnaissance d'un expert, la causalité est inverse : ce n'est pas parce qu'un membre est jugé expert qu'on écoute ses arguments, c'est parce qu'il a des arguments (et des bons) qu'il est jugé expert (voir Littlepage & Mueller, 1997, qui montre que le fait d'argumenter est un indice capital pour la reconnaissance de l'expertise dans ce type de tâche). Enfin, dans le cas de la tâche de sélection par exemple, il arrive qu'aucun membre du groupe ne possède entièrement la bonne réponse initialement mais que le groupe parvienne néanmoins à trouver la solution au problème (Moshman & Geil, 1998; Trognon, 1993). Dans ce cas, différents participants ont chacun une partie de la solution. Car ils sont défendus par de meilleures raisons, ces morceaux de solution sont adoptés par l'ensemble des participants, au contraire des autres éléments de réponse – incorrects – défendus par les mêmes personnes. Dans un tel cas il n'est pas possible de se fier à un individu : il faut que les participants soient capables d'examiner chaque élément de la réponse (et ses justifications) sur son mérite propre. Ce phénomène est à la base de *l'effet de bonus de groupe* ('assembly bonus effect'). Ce terme décrit les cas dans lesquels la performance des groupes est supérieure à celle de leurs meilleurs membres (Blinder & Morgan, 2000; Laughlin et al., 2002; Laughlin et al., 2006; Laughlin et al., 2003; Lombardelli, Proudman, & Talbot, 2005; Michaelsen, Watson, & Black, 1989; Sniezek & Henry, 1989; Stasson et al., 1991; Tindale & Sheffey, 2002). Un tel phénomène a également été observé chez les enfants, chez qui il est nommé le « deux erreurs font une bonne réponse » ('two wrongs make a right') (Glachan & Light, 1982; B. B. Schwarz, Neuman, & Biezuner, 2000).

Dans un cadre évolutionniste, on peut dire que les améliorations – parfois spectaculaires – dans ces contextes argumentatifs suggèrent que le raisonnement est *fait* pour être utilisé dans de tels contextes. On pourrait cependant rétorquer qu'il ne s'agit que d'un phénomène général de motivation : en groupe, les participants seraient plus motivés pour trouver la réponse correcte, ils passeraient plus de temps à réfléchir, feraient plus d'efforts et seraient donc plus à même de parvenir à la bonne

solution (un autre argument avancé par Oaksford et al., 1999). En d'autres termes, les participants pourraient choisir (ou non) de s'engager dans des raisonnements coûteux selon les circonstances[35]. Si cette hypothèse était exacte, elle devrait s'étendre à d'autres sources de motivations : si les groupes ont un pouvoir motivationnel particulier, il ne s'agit alors plus d'une motivation générique. Plusieurs éléments laissent penser que la motivation n'est pas ce qui fait défaut aux participants. La plupart des expérimentalistes partagerons mon intuition que, dans les tâches de raisonnement, une majorité de participants est fortement motivée (au moins pour les premiers problèmes lorsque ceux-ci sont trop nombreux). Ils demandent souvent beaucoup de temps pour répondre, voulant parfois dépasser le temps imparti qui est pourtant déjà long. Ils sont anxieux de savoir s'ils ont répondu correctement – pensant souvent qu'il s'agissait d'une forme de test d'intelligence. Plus formellement, on peut noter que dans les expériences concernant la tâche de sélection de Wason, le temps imparti aux participants pour résoudre la tâche individuellement (généralement au moins 15 minutes) leur laisse largement le temps de parvenir à une réponse dont ils soient sûrs et qu'une autre heure de réflexion ne modifierait pas.

Enfin, et ces résultats sont particulièrement intéressants, des promesses de gain ne semblent pas affecter les performances des participants. C'est la conclusion générale des études portant sur les incitations monétaires dans les tâches de raisonnement et de prise de décision (Ariely, Gneezy, Loewenstein, & Mazar, In Press; S. E. Bonner, Hastie, Sprinkle, & Young, 2000; S. E. Bonner & Sprinkle, 2002; Camerer & Hogarth, 1999). On peut prendre appui sur une expérience de Santamaria et Johnson-Laird (rapportée dans Johnson-Laird & Byrne, 2002) qui fournit un point de comparaison direct. Dans cette étude, une récompense pour la réponse correcte à la tâche de sélection de Wason était promise à certains participants. Aucune différence de performance ne fut observée entre ces participants et ceux à qui rien n'avait été promis (voir Jones & Sugden, 2001, pour un résultat similaire). La motivation pour raisonner que l'on observe dans les contextes

[35] « Findings of this kind have led to the suggestions that reasoning is by default inductive or probabilistic and that explicit deductive reasoning occurs only when people make a conscious strategic effort under instructions » (Evans, 2002, p. 986) « People may choose to engage in effortful analytic thinking because they are inclined to do so by strong deductive reasoning instructions » (Evans, 2006). Voir la section 7.8 pour d'autres réfutations de cette hypothèse.

argumentatifs est donc bien spécifique : il ne s'agit pas simplement de facteurs motivationnels génériques.

La théorie argumentative rend compte des échecs dans les tâches de raisonnement classiques, mais d'autres théories ont déjà tenté d'expliquer le fait que les participants utilisent peu ou mal le raisonnement dans ces tâches. L'explication qui est la plus souvent avancée tient aux limites de la mémoire de travail : ces limites sont censées empêcher les participants de raisonner correctement. Dans une perspective évolutionniste, cette explication est particulièrement peu convaincante. Si le raisonnement avait la fonction qui lui est habituellement prêtée, à savoir nous permettre de suivre des normes logiques afin d'améliorer la qualité épistémique de nos croyances, alors on voit mal pourquoi la sélection naturelle n'aurait pu étendre les limites de la mémoire de travail. D'un point de vue computationnel, les opérations qui sont requises pour résoudre les tâches de raisonnement sont totalement triviales par rapport à celles que doivent surmonter d'autres systèmes cognitifs. Il est ainsi très aisé de créer un programme qui résolve les syllogismes (Geurts, 2003) ou les inférences conditionnelles ou disjonctives (Rips, 1994), alors que les systèmes artificiels de vision ou de motricité sont incroyablement primaires par rapport à leurs homologues cérébraux. Il est possible de faire une analogie avec un autre domaine souffrant des limites de notre mémoire, celui du rappel de listes. Depuis Miller, on sait qu'on ne peut retenir en moyenne que sept (plus ou moins deux) éléments en mémoire à court terme. A l'inverse, la mémoire vive d'un ordinateur peut maintenant stocker des milliards de bits d'information. Une conclusion qui découle directement de ces deux éléments est que la mémoire humaine *n'est pas faite* pour mémoriser des chaînes de symboles sans signification : si elle l'était, elle n'aurait pas ces limites car elles ne peuvent être causées par de simples contraintes computationnelles[36]. De même, si la mémoire de travail contraint aussi fortement l'utilisation du raisonnement dans les tâches classiques, on peut en conclure que le raisonnement *n'est pas fait* pour résoudre ces tâches. Expliquer les échecs dans les contextes normaux des expériences de psychologie du raisonnement par les limites de la mémoire de travail ne fait que repousser le problème : il faudrait ensuite se demander

[36] De plus, d'autres formes de mémoire ont des capacités de stockage réellement gigantesques (tout au moins par rapport à la mémoire à court terme). Voir par exemple Brady, Konkle, Alvarez et Oliva (2008).

pourquoi nous avons ces limites, et à cette question on peut répondre en disant que ces limites n'ont pas été repoussées car elles ne nous gênent pas pour utiliser le raisonnement dans des circonstances naturelles, à savoir des contextes argumentatifs.

5.1.3 Pourquoi ça ne fonctionne pas tout le temps

Les améliorations sont-elles des anomalies ?

Un lecteur familier de la littérature en psychologie sociale pourrait être surpris par les effets bénéfiques qui sont attribués ici aux groupes sur le raisonnement. Dans leur récente revue de la littérature sur la prise de décision en groupe, Kerr et Tindale notent que

> The ubiquitous finding across many decades of research (e.g., see Hill, 1982; Steiner, 1972) is that groups usually fall short of reasonable potential productivity baselines—in Steiner's terminology, they exhibit process losses. (Kerr & Tindale, 2004, p.625).

Ces performances sous-optimales sont souvent expliquées par une diminution de la motivation ('group motivation loss' Steiner, 1972, ou 'social loafing' Latane, Williams, & Harkins, 1979). Cette baisse de motivation, à son tour, a été expliquée par de nombreux facteurs, mais il n'est pas pertinent de les détailler ici. Il est important de noter, par contre, que les expériences utilisées par ces courants de recherches ne sont pas des expériences qui impliquent, en général, le raisonnement.

Dans leur revue de la littérature sur le social loafing, Karau et Williams notent que

> [In the] first experiment to suggest a possible decrement in individual motivation as a result of working in a group […] male volunteers were asked to pull on a rope, tug-of-war fashion, as hard as they could in groups of varying sizes (Karau & Williams, 1993, p.682).

Ce type de tâche n'implique pas le raisonnement, ou la résolution de conflit par le biais d'arguments. La théorie argumentative ne fait donc aucune prédiction directe sur les effets que devrait alors avoir le fait de travailler en groupe. Par la suite,

> nearly 80 studies on social loafing have been conducted in which individuals' coactive efforts were compared with individuals' collective efforts. These studies have used a wide variety of tasks, including physical tasks (e.g., shouting, rope-pulling, and swimming), cognitive tasks (e.g., generating ideas), evaluative tasks (e.g., quality ratings of poems, editorials, and clinical therapists), and perceptual tasks (e.g., maze performance and vigilance tasks on a computer screen) (Karau & Williams, 1993, p.682).

Aucune de ces tâches n'est intellective au sens mentionné plus haut. Certaines tâches sont catégorisées par les auteurs comme étant 'cognitives', mais il s'agit de tâches de brainstorming. Or, ces tâches n'impliquent pas nécessairement de raisonnement : il s'agit de générer des idées, pas des arguments pour défendre ces idées. Il n'y a donc, à nouveau, pas de raison que la théorie argumentative s'y applique directement.

La polarisation des groupes

Des résultats décevants ont également été observés dans des groupes devant accomplir des tâches mettant en jeu le raisonnement. Ainsi, on a souvent observé que les discussions pouvaient avoir un effet polarisateur sur les attitudes :

> In a striking empirical regularity, deliberation tends to move groups, and the individuals who compose them, toward a more extreme point in the direction indicated by their own predeliberation judgments (Sunstein, 2002).

Que les débats aient un tel effet pourrait poser un problème pour la théorie argumentative :

> If deliberation simply pushes a group toward a more extreme point in the direction of its original tendency, do we have any systematic reason to think that discussion is producing improvements? (Sunstein, 2002)

Il a été démontré que dans certaines circonstances ce déplacement vers une opinion plus extrême est clairement non normatif (voir Bem, Wallach, & Kogan, 1965, pour la démonstration précoce d'un tel effet, et Hinsz, Tindale, & Nagao, 2008, pour une revue récente et de nouvelles expériences). Or pour la théorie argumentative les discussions en groupe forment le contexte normal du raisonnement, dans lequel il devrait produire des effets positifs. Si ces effets de polarisation de groupe étaient aussi prévalent que Sunstein l'affirme, il pourrait être difficile pour la théorie argumentative d'en rendre compte.

Il faut néanmoins souligner que cette généralisation (a 'striking empirical regularity') est quelque peu trompeuse. En réalité, pendant la plus grande partie du XX^{ème} siècle, c'est l'opinion contraire qui a prévalu : « que les jugements des groupes soient moins extrêmes que les jugements individuels est une opinion largement partagée » (Moscovici & Zavalloni, 1969, p.125). Il convient donc de s'intéresser aux circonstances qui promeuvent ce phénomène de polarisation de groupe, plutôt que de le prendre comme une conséquence automatique des discussions. Une première condition qui semble nécessaire est la présence de débats. Au début du XXème siècle, les premières expériences montrant un effet de déplacement vers un jugement moyen utilisaient des sujets sur lesquels il n'est pas facile de débattre et d'argumenter : les participants devaient estimer le poids de divers objets ou le caractère plaisant de certaines odeurs (Allport, 1924; Farnsworth & Behner, 1931).

Si la présence de débats semble nécessaire pour observer des effets de polarisation, elle n'est pas suffisante : plusieurs expériences on démontré des effets de *dé*polarisation de groupe alors même que les participants pouvaient débattre (voir par exemple Kogan & Wallach, 1966; Vinokur & Burnstein, 1978). On peut donc s'interroger sur les traits qui font que certains débats auront des effets de polarisation et d'autres des effets inverses. La théorie qui rend le mieux compte de ces effets est la théorie des arguments persuasifs ('Persuasive Argument Theory', introduite par Aviram Vinokur [1971], voir Isenberg, 1986, pour une revue, et voir Barber, Heath, & Odean, 2003, pour une application plus récente).

Selon cette théorie, les changements d'opinions résultant de discussions en groupe sont dus à l'utilisation par les participants d'arguments plus ou moins persuasifs. Dans plusieurs expériences une très forte corrélation a été observée entre la présence et la force d'arguments et la direction (et l'ampleur) du changement d'opinion s'opérant dans le groupe (Ebbesen & Bowers, 1974; Madsen, 1978). Mieux encore, d'autres expériences ont manipulé l'accessibilité d'arguments pour et contre, et on ainsi réussi à influencer les changements d'opinion du groupe, démontrant ainsi le rôle causal de la présence des arguments (voir par exemple, Kaplan & Miller, 1977, et voir Isenberg, 1986, pour de nombreuses autres références). Il est dès lors possible de rendre compte des effets de polarisation et de dépolarisation par la présence d'arguments allant soit dans le sens des attitudes initiales des membres du groupe, soit dans le sens opposé. Dans un groupe dans lequel les arguments pour et contre sont approximativement également représentés et également forts, l'opinion se déplacera vers le centre et non vers un extrême (voir par exemple Vinokur & Burnstein, 1978).

On peut alors s'interroger sur l'artificialité des effets de polarisation de groupe. Il semble en effet que pour qu'une telle polarisation soit observée, les membres du groupe doivent partager une opinion sur un sujet donné. Mais de tels contextes ne sont pas naturellement propices à l'éclosion d'un débat : si tout le monde est d'accord, il n'est pas nécessaire de débattre ou d'avancer des arguments. Une interprétation possible de la prépondérance des effets de polarisation est donc la suivante : dans des contextes expérimentaux, les participants discutent souvent de sujets qui ne prêtent pas naturellement à débat. Dès lors, le raisonnement n'est pas utilisé dans son contexte normal et il n'est donc pas si étonnant qu'il ait des conséquences épistémiquement néfastes. Il faut cependant souligner que ces contextes, bien qu'artificiels, sont loin de n'être présents que dans le cadre expérimental. On retrouve dans les institutions modernes de nombreuses circonstances qui poussent des gens qui sont en accord à discuter néanmoins – car c'est ainsi que les choses se font (les jurys), car elles doivent rendre des comptes (de nombreux comités), ou encore car leur rôle est précisément de préparer des arguments s'adressant à des personnes qui ne sont pas présentes (les responsables des campagnes politiques).

Un tel contexte est en effet celui des jurys, qui doivent discuter de certains aspects du cas même s'ils sont globalement en accord. Cela peut par exemple les

entraîner à verser des dommages beaucoup plus importants que ce que les jurés auraient versé individuellement, à condition qu'ils soient en accord sur la culpabilité de l'accusé (Schkade, Sunstein, & Kahneman, 2000). Plusieurs autres contextes, politiques, judicaires ou financiers sont discutés dans Why Societies Need Dissent (Sunstein, 2003). Mais étant donné que ces contextes requièrent un important soutien institutionnel, on peut raisonnablement penser qu'ils étaient beaucoup moins fréquents dans l'environnement dans lequel nos capacités de raisonnement ont évoluées, et donc que le problème de la polarisation ne se posait pas alors avec la force qu'il peut avoir dans notre monde moderne.

Bons conflits, mauvais conflits

Nous venons de voir que les conflits sont essentiels pour que le raisonnement en groupe produise des effets positifs. Tous les conflits, cependant, ne favorisent pas un bon fonctionnement du raisonnement en groupe. Dans le scénario proposé pour l'évolution du raisonnement, un contexte particulièrement propice à l'utilisation du raisonnement est celui des actions collectives : lorsque plusieurs personnes qui sont disposées à s'engager dans une activité commune sont en désaccord sur, par exemple, la façon de mener à bien cette activité. Les individus vont alors donner des raisons pour expliquer leurs positions respectives et ainsi essayer de convaincre les autres de l'adopter, tout en comprenant et prenant en compte leurs propres arguments. Il est important que les individus qui débattent partagent des intérêts communs, des valeurs communes, qu'ils se fassent généralement confiance. La littérature sur le rôle des conflits et de la confiance au sein des groupes, et en particulier au sein des organisations, est immense, et il n'est pas question de la passer ici en revue. Je me contenterai de citer quelques études qui sont représentatives de l'ensemble et qui illustrent parfaitement le rôle que peuvent jouer différents types de conflit.

La distinction entre deux types de conflits a tout d'abord été opérationnalisée dans un article de 1954 distinguant les conflits basés sur la substance de la tâche et ceux basés sur les relations interpersonnelles entre membres du groupe (Guetzkow & Gyr, 1954). Reprenant Jehn (1995), Simons et Peterson résument ainsi la difference entre ces deux types de conflits :

> Task conflict, or cognitive conflict, is a perception of disagreements among group members about the content of their decisions and involves differences in viewpoints, ideas, and opinions. Relationship conflict, or emotional conflict, is a perception of interpersonal incompatibility and typically includes tension, annoyance, and animosity among group members (Simons & Peterson, 2000, p.102).

De nombreuses études convergent vers le résultat suivant : les conflits portant sur la tâche ont généralement des effets positifs sur les performances alors que les conflits relationnels ont des effets négatifs (voir Simons & Peterson, 2000, et les nombreuses références qu'ils citent).

Ces résultats s'accordent bien avec les prédictions de la théorie argumentative pour laquelle un accord préalable (généralement tacite) sur la possibilité de coopérer est nécessaire – ce qui se traduit ici par une absence de conflits relationnels. Dès lors que cet accord est acquis, les bénéfices du raisonnement s'obtiennent principalement lorsqu'il y a conflit sur la façon d'opérer – les conflits liés à la tâche (sinon, il n'y a pas de raison d'argumenter et donc de raisonner).

On retrouve une dichotomie très similaire dans une étude de grande ampleur de Jehn et ses collègues, qui ont analysé les comportements en groupe d'employés d'une grande entreprise américaine (Jehn, Northcraft, & Neale, 1999). Les différences entre membres des groupes étaient synthétisées en trois catégories : différences sociales (genre, ethnicité), différences de valeur (portant sur l'objectif à accomplir), et différences informationnelles (différences d'éducation et d'expertise). Si on s'intéresse aux performances objectives des groupes, on observe que les différences de valeurs étaient corrélées à de plus mauvaises performances alors qu'au contraire les différences informationnelles étaient elles corrélées à de meilleures performances – ce trait étant renforcé dans le cas des tâches complexes[37]. Des effets similaires ont été reportés par Jehn (1995) et par Jehn et Mannix (2001). A nouveau, ces résultats s'accordent avec la théorie argumentative : les désaccords sur l'objectif général ont des effets négatifs alors que ceux portant sur la façon de l'accomplir ont des effets positifs.

[37] Les différences sociales n'ayant aucun effet.

Les résultats qui ont été examinés ici corroborent les prédictions de la théorie argumentative : lorsque des participants sont en désaccord sur la façon de résoudre un problème, et qu'ils en discutent en groupe, ils sont capables à la fois de produire des arguments pour défendre leur point de vue, et d'évaluer ceux des autres, de telle façon que la solution la mieux défendue l'emporte – et il s'agit généralement de la bonne solution dans les tâches de raisonnement. Il est très difficile de rendre compte de ces résultats par d'autres facteurs qu'une utilisation supérieure du raisonnement dans ces contextes par rapport aux contextes de résolution individuelle. Bien que ce type d'amélioration puisse faire figure d'exception parmi les performances des groupes plus généralement, ceci n'est pas un problème pour la théorie argumentative. Au contraire, cette dernière n'a pas de raison de s'appliquer à la majorité des tâches de groupe (qui n'impliquent pas le raisonnement), et dans les cas ou le raisonnement peut être responsable de mauvaises performances en groupe, une théorie faisant les mêmes prédictions que la théorie argumentative – la théorie des arguments persuasifs – peut rendre compte des résultats. Finalement, les conclusions des études sur le type de conflit qui peut servir de catalyseur ou au contraire handicaper un groupe sont également conformes aux prédictions de la théorie argumentative.

5.2 L'argumentation

Si les contextes discursifs sont idéaux pour activer le raisonnement, ils posent des problèmes méthodologiques : il est très difficile d'analyser ce qui se passe dans ces groupes. S'il est difficile d'expliquer la convergence vers la bonne réponse autrement que par de bonnes capacités de production et d'évaluation d'arguments, il n'en reste pas moins que ces études nous renseignent mal sur la façon dont fonctionnent ces capacités. C'est pourquoi nous nous tournons maintenant vers des recherches ayant tenté de répondre plus précisément à ces questions. Du point de vue de la théorie argumentative, on ne peut pas vraiment s'attendre à d'aussi bonnes performances que dans le cas des groupes : les contextes utilisés ici auront presque tous un côté un tant soit peu artificiel. Il n'empêche que, dans la mesure où la situation se rapproche d'une situation d'argumentation, les performances devraient

être satisfaisantes – supérieures, en tout cas, aux performances face à des problèmes de raisonnement plus classiques.

5.2.1 Compréhension et évaluation d'arguments

La psychologie sociale : persuasion et changement d'attitude

La façon dont les gens comprennent et évaluent différents arguments a été principalement étudiée dans deux domaines de recherche : la psychologie du raisonnement et une branche de la psychologie sociale portant sur la persuasion et le changement d'attitude. C'est par cette dernière que nous allons commencer.

L'étude de la persuasion et du changement d'attitude a une longue histoire. On pourrait bien entendu considérer qu'elle commence avec les rhéteurs grecs, mais si on se cantonne à la psychologie 'scientifique', cette discipline prend son envol pendant la seconde guerre mondiale, pour répondre aux besoins de propagande du gouvernement américain (voir Billig, 1996, pour une tentative d'intégration des vieux courants rhétoriques avec la psychologie moderne). Les chercheurs vont alors étudier l'impact de messages persuasifs en fonction de trois grandes catégories de facteurs : ceux liés à la source du message, ceux liés au récepteur, et ceux liés au message lui-même. Malgré des avancées certaines, les expériences donnent souvent des résultats apparemment contradictoires, et il faudra attendre la fin des années 70 pour qu'émergent des théories qui parviennent à rendre compte de ces découvertes. Les théories dominantes sont toutes deux des théories à processus duel, assez proches l'une de l'autre, celle de la probabilité d'élaboration ('elaboration likelilhood model' [Petty & Cacioppo, 1986]) et le modèle heuristique-systématique (Chaiken, Liberman, & Eagly, 1989). Selon ces théories, les gens peuvent réagir de deux façons différentes face à un message persuasif. S'ils ne sont pas motivés, ils utiliseront des indices périphériques, ou heuristiques, pour déterminer la force du message : quel est le statut du locuteur, est-il attrayant, fait-il partie de mon groupe, etc. Ils ne prêtent alors pas vraiment attention aux arguments eux-mêmes, et ne discriminent donc pas entre des arguments forts et faibles. Si le public est motivé, au contraire, il prendra justement en compte la force des arguments et non plus les indices périphériques (ou

ceux-ci joueront un rôle moindre) (voir Petty, Cacioppo, & Goldman, 1981 et Crano & Prislin, 2006; Petty & Wegener, 1998, pour revues). Il s'agit là d'un résultat important, et qui contraste fortement avec ceux obtenus en psychologie du raisonnement : lorsque les conclusions sont pertinentes, les participants sont parfaitement à même d'accorder plus de poids aux arguments forts qu'aux faibles.

Dans la mesure où les gens sont souvent influencés par des facteurs qui ne devraient pas être pris en compte (tels que le charme de différentes personnes), on pourrait tirer de ces résultats un constat d'échec. Mais si on considère l'activité d'évaluation d'arguments dans un contexte plus large, on peut au contraire les interpréter comme reflétant une bonne stratégie de régulation des efforts. Lorsque nous n'accordons qu'une importance très limitée au contenu du message, il semble raisonnable de ne pas dépenser beaucoup d'énergie pour l'évaluer (nous avons d'autres choses auxquelles penser après tout). Par contre, cette littérature montre clairement que, dans les bonnes circonstances, lorsqu'ils sont motivés et ne sont pas distraits, les gens sont parfaitement capables d'évaluer des arguments selon leur force, et d'être influencés en fonction d'elle. L'étude de la persuasion et du changement d'attitude fournit donc beaucoup d'informations sur les contextes dans lesquels les gens sont disposés à utiliser leurs capacités de raisonnement – et non d'autres processus heuristique – pour évaluer des arguments. Elle ne nous donne, cependant, presque aucune information sur la façon dont ces arguments sont évalués, comme le notent (et l'avouent) deux des chercheurs dominant cette discipline : « relatively little is known about what makes an argument persuasive. » (Petty & Wegener, 1998).

Psychologie du raisonnement

Alors que des psychologues sociaux s'intéressaient aux facteurs qui promeuvent ou non l'utilisation des capacités de raisonnement, les psychologues du raisonnement s'intéressaient au fonctionnement de ces mêmes capacités. Pour ce faire, un type de tâche à laquelle ils ont souvent recours consiste à demander aux participants d'évaluer des arguments. Il est cependant important de souligner que ces tâches n'ont que peu de rapport avec celles qui sont utilisées en psychologie sociale. Si dans les deux cas les participants doivent bien évaluer des arguments, le type

d'argument à évaluer, et ce qui est demandé, est très différent. En psychologie sociale, il s'agit d'arguments réalistes (par exemple pour ou contre la peine de mort), et on ne demande pas aux participants de les évaluer directement : on mesure l'évolution de leur attitude concernant le domaine visé par les arguments. En psychologie du raisonnement, les arguments sont presque toujours artificiels (tous les bateleurs sont des pianistes, tous les pianistes sont des boulangers, donc tous les bateleurs sont des boulangers), et on demande aux participants de les évaluer par rapport au strict étalon de la logique : ils doivent se prononcer sur la validité logique des arguments présentés.

On ne peut donc guère comparer les performances dans les deux types de tâche. Il est cependant intéressant de noter que, à l'aune des résultats de la psychologie sociale, on pourrait prédire que les participants des expériences de psychologie du raisonnement ne devraient, en fait, que peu raisonner (!). En effet de nombreux travaux de psychologie sociale indiquent que « le facteur le plus important déterminant la motivation d'une personne à réfléchir est la pertinence personnelle de la communication » (Petty & Wegener, 1998, voir Johnson & Eagly, 1989; Petty & Cacioppo, 1979, 1990). Or les arguments utilisés en psychologie du raisonnement n'ont généralement *aucune pertinence* pour les participants (nous verrons par la suite qu'il existe quelques exceptions). Les psychologues du raisonnement ont commencé d'utiliser ce type d'arguments car ils souhaitaient justement faire abstraction des effets de contenu : ils ne voulaient pas savoir si les gens étaient capables de raisonner sur la peine de mort, ou l'avortement, mais s'ils étaient capables de raisonner en général. Par ailleurs, la théorie poppérienne qui était alors dominante en philosophie des sciences, et qui a eu une influence importante sur certains fondateurs de la psychologie du raisonnement (on pense à Peter Wason), mettait en avant des capacités très générales, de falsification d'hypothèses en particulier.

Quelqu'en soient les raisons exactes, il se trouve que les psychologues du raisonnement utilisent le plus souvent des arguments pour lesquels il est maintenant bien connu que les gens tendent non pas à raisonner, mais à utiliser des mécanismes heuristiques. Il s'agit là probablement de la cause principale des mauvaises performances observées dans les tâches de psychologie du raisonnement. Ces considérations rejoignent les prédictions de la théorie argumentative : pour elle, les contextes des tâches de raisonnement, qui n'ont rien d'argumentatif, n'engagent pas naturellement nos capacités de raisonnement (rappelons que ces mêmes problèmes

dans un contexte argumentatif – débat en groupe – peuvent être résolus par un très grand nombre de participants – cf. section 5.1.1).

Etant donné le contexte des expériences classiques de psychologie du raisonnement, on peut essayer de faire des prédictions sur les heuristiques dont elles devraient favoriser l'utilisation. On peut tout d'abord s'interroger sur la réaction initiale des participants. Les résultats de Gilbert, mentionnés dans la section 4.4 sont pertinents ici ; voici la conclusion qu'en avaient tirés Petty et Wegener : « in the absence of contrary information, people tend to assume that what others say is true » (Petty & Wegener, 1998). Parmi ces informations allant à l'encontre de ce qui est énoncé, j'avais mentionné la présence de croyances contradictoires en mémoire et le fait que la source soit notoirement malveillante ou incompétente. Dans le cas des arguments pour lesquels la conclusion doit être évaluée, il faut ajouter aux croyances des participants les éléments donnés dans les prémisses. Etant donné que les participants n'ont pas de raison de se méfier de la source des informations dans une expérience de psychologie du raisonnement, et que l'utilisation de contenus artificiels empêche toute interférence des croyances préalables, la première réaction devrait être dirigée uniquement par les prémisses présentées : s'il n'y a pas d'incohérence, la première réaction devrait être une réaction d'acceptation. Il est possible que les participants essaient ensuite de remettre en cause cette réaction initiale, mais étant donné qu'ils ne devraient pas naturellement être motivés à examiner de façon critique l'argument, rien ne garantit qu'ils le fassent.

On peut se contenter ici de citer les résultats de deux études qui illustrent ce phénomène, au moyen des deux types de problèmes les plus courants en psychologie du raisonnement : les inférences conditionnelles et les syllogismes classiques (basés sur les quantificateurs). La première est une étude de grande ampleur portant sur l'intégralité des syllogismes possibles (256 combinaisons) (Evans, Handley, Harper, & Johnson-Laird, 1999). Malgré cette grande diversité de problèmes, les résultats s'interprètent assez simplement dans le cadre de la théorie des modèles mentaux. Cette théorie postule que les participants créent un modèle mental simplifié des prémisses auquel ils comparent celui de la conclusion. Si le modèle de la conclusion ne correspond pas à celui des prémisses, ils peuvent la rejeter. Sinon, ils commencent par accepter la conclusion, en se basant sur ce modèle simplifié, puis doivent former et parcourir un modèle plus complet à la recherche de contre exemples afin de s'assurer que la conclusion est bien valide. Les résultats de cette étude ont montré

que les participants ne cherchaient en fait pas de contre exemples, et se contentaient d'accepter la conclusion si elle était contenue dans le modèle simplifié (Evans et al., 1999, p.1502). Dans ce cas une simple heuristique consistant à accepter ce qui est cohérent avec nos croyances (dans ce cas les prémisses que nous venons d'accepter) et à rejeter le reste suffit à rendre compte des résultats (ces résultats seront examinés plus en détail dans la section 6.3).

La seconde illustration est fournie par une étude dans laquelle les participants devaient évaluer des inférences conditionnelles (si p alors q, p, donc q ?). Dans une condition, les instructions mettaient l'accent sur le temps de réponse, et demandaient aux participants de répondre aussi vite que possible (Schroyens, Schaeken, & Handley, 2003). On observe alors que les participants acceptent massivement les quatre inférences conditionnelles[38] (à plus de 80%), ce qui peut s'expliquer par l'utilisation d'une heuristique d'acceptation en l'absence d'éléments contraires. Nous verrons plus bas que les résultats des études sur le biais de croyance pointent dans la même direction (voir section 6.3).

Certains participants, bien entendu, ne se contentent pas d'utiliser ces heuristiques et cherchent plus avant la solution correcte. Le point important ici est que les mauvaises performances des participants dans les tâches d'évaluation d'arguments utilisées en psychologie du raisonnement ne sont pas le reflet d'une mauvaise capacité à évaluer des arguments, mais simplement un manque de motivation.

Tâches classiques en contexte argumentatif

En marge de la psychologie du raisonnement s'est récemment développé un domaine restreint d'étude de l'argumentation. Il peut s'agir de recherches directement inspirées de la psychologie du raisonnement : par exemple l'étude de la compréhension d'énoncés conditionnels dans des contextes argumentatifs. D'autres expériences portent sur des traits propres à l'argumentation, tels que la charge de la preuve, l'engagement, ou encore les paralogismes de l'argumentation. Dans tous les

[38] Si p alors q, puis : p, donc q (modus ponens, valide) ; q, donc p (affirmation du conséquent, invalide) ; non-q, donc non-p (modus tollens, valide) ; non-p, donc non q (déni de l'antécédent, invalide).

cas, les arguments utilisés se rapprochent des arguments que l'on peut rencontrer dans la vie de tous les jours. En voici un exemple :

> Barbara: Are you taking digesterole for it?
> Adam: Yes, why?
> Barbara: Well, because I strongly believe that it has side effects.
> Adam: It doesn't have any side effects.
> Barbara: How do you know?
> Adam: Because I know of an experiment where they failed to find any
> (tiré de Oaksford & Hahn, 2004, p.84)

Le contraste est net avec les stimuli habituels des expériences de psychologie du raisonnement : on comprend aisément que ce type de problème engage plus facilement l'attention des participants qui rencontrent quotidiennement ce type d'interaction. Par ailleurs, les compétences testées dans ces études sont indispensables à qui veut être capable de débattre ou même de suivre un débat entre d'autres personnes. Etant donné que les stimuli utilisés sont proches des stimuli que ces capacités sont faites pour traiter, on devrait s'attendre à de bonnes performances.

Avant de se tourner vers les études portant sur des traits spécifiquement argumentatifs, on peut mentionner une expérience plus proche de la psychologie du raisonnement classique. Thompson et ses collègues ont confronté les participants à des conditionnels ayant une visée argumentative, tels que « if the Kyoto accord is ratified, there will be a downturn in the economy », conditionnels qui étaient insérés dans le contexte plus large d'un article de journal (V.A. Thompson, Evans, & Handley, 2005). Ils ont ensuite testé la compréhension des intentions de la personne ayant proposé l'argument, en demandant par exemple « est-ce que le locuteur est favorable aux accords de Kyoto ? ». Dans ce cas, les participants n'avaient aucun problème à comprendre le point de vue du locuteur. Le raisonnement nécessaire pour tirer une telle conclusion est proche d'un *modus tollens* (si p alors q, non q, donc non p) : les participants doivent comprendre qu'un ralentissement de l'économie est une conséquence négative, que si les accords de Kyoto entraînent une conséquence négative, il s'agit d'une raison de les rejeter, et donc que le locuteur est opposé à ces accords. La très bonne performance des participants dans cette tâche est intéressante

car elle contraste avec les très mauvaises performances observées lorsque les participants doivent évaluer des *modus tollens* artificiels (auquel cas on observe généralement 40% d'erreur pour un choix dichotomique). Cette étude indique donc que lorsqu'un argument a un contenu argumentatif approprié, il est compris beaucoup plus facilement.

Les paralogismes de l'argumentation

On peut distinguer deux grands thèmes dans les études portant plus spécifiquement sur l'argumentation : la structure globale des arguments et les paralogismes. L'étude des paralogismes de l'argumentation se rapprochant d'avantage de l'étude des autres erreurs de raisonnement, c'est par elle que nous allons commencer.

Un groupe de chercheurs Israéliens a été parmi les premiers à étudier la capacité à repérer les paralogismes de l'argumentation. Le paradigme général qu'ils ont utilisé est le suivant : ils présentent deux personnages, disent qu'ils sont en train de débattre d'un sujet donné, et donnent un des arguments utilisés, qui peut-être un paralogisme. On demande aux participants s'ils pensent que l'argument utilisé est bon ou s'il y a au contraire un problème. D'autres tâches sont également utilisées mais elles sont plus formelles (reconnaissance des normes d'argumentation sous leur forme abstraite par exemple), et on peut donc se contenter de rapporter les résultats de la tâche de détection des paralogismes. Plusieurs expériences ont été menées à bien, portant sur les paralogismes les plus connus : *ad hominem* (attaque contre la personne), *ad populum* (argument se basant sur l'opinion de la majorité), *ad ignorantiam* (argument du type 'Dieu existe car on n'a pas prouvé qu'il n'existe pas'), ou encore l'argument de la fausse cause (*post hoc ergo propter hoc* : inférer la causation à partir de la contiguïté temporelle). Dans l'ensemble, les participants ont de bonnes performances : respectivement 75, 72 et 62% d'identifications correctes (Neuman, 2003; Neuman, Weinstock, & Glasner, 2006; Weinstock, Neuman, & Tabak, 2004). Ces résultats sont d'autant plus intéressants que les participants étaient des collégiens ou des lycéens, âgés de 15 ou 16 ans en moyenne. Or on sait qu'à cet âge les participants ont tendance à se reposer davantage sur l'heuristique d'acceptation, et donc à ne rejeter que très peu de conclusions pour ce qui est des

tâches classiques de psychologie du raisonnement (voir par exemple Barrouillet, Grosset, & Lecas, 2000, sur les performances à une tâche d'inférence conditionnelle par des participants de 15 ans). Il est donc frappant que des participants de cet âge soient capables de détecter la faiblesse de ces arguments alors qu'ils ne peuvent déterminer quelles inférences conditionnelles sont invalides.

Le rôle du contexte argumentatif est un des facteurs modérateurs étudiés par ces chercheurs. Ceci est pertinent ici car la théorie argumentative prédit que les performances devraient se détériorer en dehors de tels contextes. C'est ce qu'ont observé Neuman et al. (2006) en comparant un dialogue argumentatif (se basant sur des raisons) et un dialogue visant simplement à blesser l'autre personne : les performances – de détection de paralogismes – étaient supérieures dans le premier cas.

Lance Rips s'est également intéressé aux paralogismes de l'argumentation, et en particulier au raisonnement circulaire (ou 'begging the question') (Rips, 2002, voir également Brem, 2003). Dans un argument circulaire, une conclusion est utilisée pour se soutenir elle-même :

> Allen: The Evanston City Council should make it illegal to tear down the city's old warehouses.
> Beth: What's the justification for preserving them?
> Allen: The warehouses are valuable architecturally.
> Beth: Why are they so valuable?
> Allen: The older buildings lend the town its distinctive character.
> Beth: But what's the reason the warehouses give it character?
> Allen: The warehouses are valuable architecturally.
> (tiré de Rips, 2002, pp.768-9)

Pour échapper à la circularité Allen aurait dû avancer un argument différent à la dernière ligne du dialogue. Rips a observé que les participants étaient capables de distinguer les arguments circulaires en les notant comme étant moins raisonnables que des arguments ne contenant pas de répétition ou dans lesquels la répétition n'est pas problématique.

Plus récemment des expériences similaires ont été menées à bien avec de très jeunes enfants. Les participants étaient confrontés à plusieurs explications pour un

problème donné, dont certaines étaient circulaires et d'autres non (Baum, Danovitch, & Keil, 2007). Dès 6 ans, les enfants préfèrent l'explication non circulaire à l'explication circulaire (ils la choisissent au dessus du niveau attendu au hasard), et à 10 ans ils sont plus de 80% à la choisir. A nouveau, il est frappant de constater que des enfants aussi jeunes soient sensibles à ce type de différences alors que leurs lacunes dans les tâches de raisonnement abstrait sont très importantes.

Bien qu'elles étudient le raisonnement informel, ces études ont partiellement hérité des normes plus strictes de la psychologie du raisonnement classique : on considère généralement que les participants se trompent si, par exemple, ils ne repèrent pas un paralogisme. Mais, de même que certains arguments qui ne sont pas logiquement valides peuvent être très bons ('sound'), certains paralogismes sont en fait de bons arguments. Or les instructions demandent justement aux participants d'identifier les arguments défaillants : si un paralogisme est un bon argument, il est dès lors normal qu'ils ne le signalent pas. Par exemple, le raisonnement circulaire qui vient d'être donné en exemple est clairement un mauvais argument, qui revient à dire que les entrepôts ont une valeur architecturale car ils ont une valeur architecturale. On peut le comparer à un des arguments *ad ignorantiam* utilisé par Weinstock et al. (2004) :

> During a lesson they debate the question: "Is it justified to sell weapons to nondemocratic countries?"
> Don argues that it is justified to sell weapons to non-democratic countries.
> Henry argues that it is not justified to sell weapons to non-democratic countries.
> During the debate Don argues: "It is justified to sell weapons to non-democratic countries because no one has proven it is not justified to behave this way."

Cet argument a une certaine force. On peut penser qu'il serait assez facile de démontrer que, si tel est le cas, la vente d'armes à des pays non démocratique a des conséquences tragiques. Et il est également raisonnable de supposer que bon nombre de chercheurs se sont penchés sur la question. S'il est vrai que personne n'a démontré qu'il n'est pas justifié de vendre des armes à des pays non démocratiques, c'est en effet un bon argument pour penser qu'il n'y a pas d'arguments factuels forts

contre ce type de vente d'arme, ce qui est un argument recevable pour penser qu'il est justifié de le faire.

Afin de pouvoir déterminer quels paralogismes sont de réels paralogismes qu'il convient de rejeter, Oaskord, Hahn et leurs collaborateurs ont construit plusieurs théories visant différents paralogismes (pente glissante [slippery slope], *ad ignorantiam* et arguments circulaires). Dans la continuité des travaux d'Oaksford et ses collègues sur le raisonnement déductif et inductif, ces études utilisent le formalisme bayésien pour modéliser la façon dont sont évalués différents arguments. Cette méthode présente plusieurs avantages. Tout d'abord, il ne s'agit pas d'une mesure dichotomique (paralogisme ou non), mais d'une échelle portant sur l'effet d'un argument sur le degré de croyance dans une proposition : un argument fort est un argument qui modifie considérablement nos croyances préalables (les priors). De plus, il est assez aisé d'intégrer diverses variables afin de modéliser différents types d'arguments : nous allons illustrer cela avec le cas de l'*ad ignorantiam*. Un exemple d'*ad ignorantiam* que l'on considère généralement valide est :

(1) Le médicament A n'est pas toxique car aucun effet toxique n'a été observé dans une expérience.

Bien que cet argument soit considéré comme valide, il est moins fort que l'argument suivant :

(2) Le médicament A est toxique car des effets toxiques ont été observés dans une expérience.

De même, le premier argument est moins fort que l'argument suivant :

(3) Le médicament A n'est pas toxique car aucun effet toxique n'a été observé dans 50 expériences.

Enfin, le degré de croyance initial peut également influer sur l'acceptation de l'argument : on sera plus facilement convaincu par (1) si on avait déjà une forte présomption que ce médicament était inoffensif (Neuman, Glassner, & Weinstock, 2004, font la même remarque).

Un cadre bayésien permet de prendre en compte ces différents facteurs. Rappelons la formule de Bayes :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}.$$

Il s'agit de définir la probabilité que A étant donné B, en fonction de la probabilité a priori de A, de son contraire (A barre), et des probabilités de B étant donné A et de B étant donné le contraire de A.

Dans le cadre de l'argument *ad ignorantiam* concernant le médicament, cette formule peut être adaptée ainsi :

$$P(T|e) = \frac{nh}{nh + (1 - l)(1 - h)}$$

(Hahn & Oaksford, 2007, p.709). Ici, T est la probabilité que le médicament soit toxique ; e (-e) une expérience dans laquelle la toxicité est (n'est pas) observée ; n la sensitivité du test P(e|T) ; l sa spécificité P(-e|-T) et h la probabilité a priori que le médicament soit toxique. A partir de cette équation, on peut montrer que pour tous les tests ayant un pouvoir discriminatoire au dessus du hasard, alors la version négative de l'argument [(1)] est moins forte que la version positive [(2)]. Il est facile, dans cette équation, d'intégrer le nombre d'expériences menées à bien, et de montrer qu'il a l'influence souhaitée. Enfin, les probabilités a priori sont encore plus aisément prises en compte car déjà intégrées dans l'équation.

Hahn et Oaksford on ensuite validé les prédictions du modèle sur la force de différents arguments. Pour cela, ils ont inclus les arguments dans un débat entre deux personnes et demandé aux participants d'indiquer si la personne a qui est adressé l'argument doit en accepter la conclusion (Oaksford & Hahn, 2004, voir également Hahn, Oaksford, & Bayindir, 2005). Les résultats montrent que les participants prennent en compte les différents facteurs mentionnés ci-dessus : ils estiment que l'interlocuteur doit être plus convaincu par les arguments positifs, par ceux soutenus par un plus grand nombre d'expériences, et ils sont influencés par sa croyance préalable dans l'efficacité du médicament.

Des modélisations, accompagnées de leur soutien expérimental, ont été menées à bien pour d'autres types d'arguments : le raisonnement circulaire (expériences 1 et 2 de Hahn & Oaksford, 2007) et la pente glissante (Corner, Hahn, & Oakfsord, 2006). Enfin, les auteurs ont procédé à une dernière validation de leur

hypothèse en comparant directement certains des arguments dont ils se sont servis dans leurs expériences, qui ont tous une certaines force, aux formes classiques des paralogismes, qui sont généralement beaucoup plus faibles (Hahn & Oaksford, 2007, expérience 3). Dans tous les cas, les participants étaient capables de les distinguer et d'attribuer moins de force aux formes classiques des paralogismes.

Ces résultats montrent que les participants ne catégorisent pas les arguments en paralogisme/argument valide. Ils les évaluent sur la base de leur force, en tenant compte de divers facteurs qui sont en effet pertinents – ce à quoi on pourrait s'attendre compte tenu du décours des débats de la vie quotidienne. La force de cette perspective tient justement au fait qu'elle permet de prendre en compte les différents traits qui font qu'un argument est fort ou faible, plutôt que de simplement se demander s'il s'agit d'un paralogisme ou non. On peut cependant s'interroger sur l'apport du formalisme bayésien. Si son caractère très général et flexible lui permet d'accommoder de nombreux types d'arguments, il n'est pas clair que les validations expérimentales lui apportent un soutien, au-delà des intuitions qu'il permet de formaliser. Dans le cas de l'argument par ignorance par exemple, les auteurs incorporent leurs intuitions sur ce qui fait qu'un argument est fort ou faible dans les formules bayésiennes. Ensuite, ils observent une très bonne corrélation entre le modèle et les performances des participants. Mais il s'agit du modèle dont plusieurs paramètres ont été ajustés pour correspondre justement aux données. Dans ce cas, le modèle n'apporte aucune précision quantitative. Or les prédictions qualitatives ont été faites sur la base des intuitions, préalablement à la modélisation. Ce que le modèle apporte n'est donc pas clair.

Dans l'ensemble, ces expériences portant sur les paralogismes de l'argumentation montrent que les participants (y compris des adolescents, et même de jeunes enfants dans un cas) sont capables de les repérer lorsqu'ils ne sont pas adéquats, et d'évaluer leur force correctement lorsqu'ils le sont.

La structure globale des arguments

Les études passées en revue jusqu'à présent continuent, dans la lignée de la psychologie du raisonnement, de porter sur des arguments individuels. Mais comprendre ou participer à un débat nécessite d'autres capacités, des capacités traitant de la structure plus globale des arguments présentés, du contexte dans lequel ils s'intègrent. Il faut ainsi pouvoir comprendre ou se situe un argument particulier dans le schéma plus général de la discussion, qui s'est engagé vis-à-vis de tel ou tel énoncé, ou encore à qui revient, à chaque moment, la charge de la preuve.

Une étude s'est penchée sur la façon dont les gens perçoivent l'articulation des prémisses et des conclusions dans un argument (Ricco, 2003). Dans l'expérience principale les participants devaient distinguer trois types de liens entre prémisses et conclusion : l'indépendance, la coordination partielle et la coordination complète. Des prémisses sont indépendantes lorsque le soutien qu'elles apportent à la conclusion est totalement indépendant : si une s'avère être fausse, cela ne modifie nullement le soutien apporté par l'autre prémisse. En voici un exemple (les exemples sont tirés de Ricco, 2003, p.1027) :

Conclusion:	The possession of marijuana for personal use should be decriminalized.
Prémisses:	(1) Marijuana can lessen pain in people with chronic pain conditions.
	(2) Marijuana can reduce nausea in patients undergoing chemotherapy.

Dans un argument coordonné, une des prémisses renforce l'autre, ou elles se renforcent mutuellement. Ce renforcement peut-être plus ou moins fort, il est complet si la fausseté d'une prémisse sape tout le soutien que l'autre apportait à la conclusion. Dans l'exemple suivant,

Conclusion:	We should stop eating so many pre-packaged, microwave dinners.
Prémisses:	(1) Microwave dinners contain a lot of salt.

(2) Eating lots of salt can cause high blood pressure.

la prémisse (2) est totalement coordonnée à (1) (si (1) est faux, (2) n'apporte plus aucun soutien à la conclusion), alors que (1) n'est que partiellement coordonné à (2) (si (2) est faux, (1) peut tout de même apporter un soutien à la conclusion – si manger trop de sel est mauvais pour d'autres raisons par exemple). Il s'avère que les participants sont tout à fait capables de juger si des prémisses sont indépendantes, ou si elles sont partiellement ou pleinement coordonnées. Dans d'autres expériences, ces conclusions ont été étendues à d'autres distinctions, plus subtiles (par exemple aux cas de subordination dans lesquelles une prémisse apporte un soutien à une autre prémisse).

Deux études ont quant à elles examiné le problème de la charge de la preuve et de l'engagement. Dans une première série d'expériences, Baillenson et Rips (1996, voir également la troisième expérience de Rips, 1998) ont présenté des extraits de débats aux participants :

1. Rita: Abortion should not be legal
2. Alan: But the right to have an abortion for a woman is fundamental
3. Rita: I am no expert on rights or anything, but I think that abortion is not an issue of freedom
4. Alan: The freedom to control your own body is one of the most important rights a person can have
5. Rita: What is your evidence for that statement?

et leur ont demandé de déterminer à quelle personne revenait la charge de la preuve (la personne qui a 'more to prove' pour reprendre leurs instructions). Ils ont étudié l'influence de trois facteurs sur ce choix. Tout d'abord l'initiation du débat : quelle personne fait la première remarque controversée. Ils ont observé que celle-ci portait, toutes choses égales par ailleurs, d'avantage la charge de la preuve. Ils ont également modulé la force de certains des arguments, et les participants y réagirent correctement, assignant d'avantage la charge de la preuve à la personne donnant les arguments les plus faibles. Enfin, ils ont fait varier l'emplacement du questionnement ('What is your evidence for that statement?') : les participants sont bien sensibles au

fait que dans certains débats (dont celui utilisé en exemple), l'interlocuteur n'y répond pas et il reçoit donc davantage la charge de la preuve.

Rips a également élaboré un modèle de la façon dont les intervenants d'un débat s'engagent, et restent ou non engagés vis-à-vis de différents énoncés (Rips, 1998). Il s'agit d'un ensemble de règles gouvernant les réactions que les participants devraient avoir face à certains types d'arguments. Par exemple, le fait d'accepter une justification entraîne l'acceptation de l'énoncé ainsi justifié. Par contre, le fait d'accepter une répartie entraîne le rejet de l'énoncé ainsi attaqué. Des participants eurent à juger si les intervenants d'un débat étaient ou non engagés vis-à-vis des énoncés qu'ils avaient prononcé à divers moments du débat. Leurs réponses indiquèrent qu'ils respectaient le modèle de Rips. Or il s'agit également d'un modèle normatif : s'il est respecté, les intervenants d'un débat sont nécessairement en accord sur qui est engagé vis-à-vis de tel ou tel énoncé. Il s'agit donc d'un élément supplémentaire montrant que des participants naïfs ont une bonne compréhension des principes de l'argumentation, compréhension nécessaire à la fois pour suivre et pour s'engager dans un débat.

Une expérience sur l'évaluation d'arguments

Une étude conduite par Keith Stanovich et Richard West fournit le cadre de celle que je vais relater ensuite (Stanovich & West, 1999). Ils ont utilisé une série de problèmes classiques en raisonnement et prise de décision, avec la modification suivante : après une première passation d'un problème de raisonnement, les participants devaient lire et évaluer un argument défendant la bonne (ou la mauvaise) réponse. Ensuite, ils étaient confrontés à d'autres tâches puis devaient résoudre le problème initial une seconde fois, alors que les instructions les incitaient à donner une réponse qui pouvait être différente de la première et pouvait prendre en compte l'argument donné. Cette méthode permet de tester l'influence d'explications sur le comportement des participants : s'ils comprennent réellement les explications fournies, ils devraient être capables de donner la bonne réponse lorsqu'ils passent le test pour la seconde fois.

Les deux premières tâches portaient sur des questions de probabilité, et plus précisément sur les éléments à prendre en compte pour faire un diagnostic : faut-il

prendre en compte le taux de base et les chances que les données soient observées alors même que l'hypothèse est fausse – ces deux éléments tendant à être sous évalués dans ce type de tâche. Les participants étaient divisés en trois conditions. Dans la première, ils recevaient un argument pour la bonne réponse, dans la seconde un argument pour la mauvaise, et dans la troisième les deux. Les résultats furent exactement ce que l'on prédit sur la base d'une bonne compréhension, évaluation et prise en compte des arguments : dans les première et troisième conditions (lorsqu'ils étaient confrontés au moins au bon argument), plus de participants ont changé de la mauvaise vers la bonne réponse que dans la direction opposée, alors que dans la seconde condition, ne comprenant que l'argument pour la mauvaise réponse, il n'y avait pas d'effet significatif de l'argument. Cela indique que le bon argument seul a une influence, qu'il conserve cette influence lorsqu'il est confronté à l'argument pour la mauvaise réponse, alors que ce dernier n'en a aucune. Des résultats similaires furent observés dans le cas du problème des coûts irrécupérables ('sunk cost'). Par contre, dans le dernier problème de raisonnement étudié (le paradoxe de Newcomb), l'argument pour la mauvaise réponse fut aussi efficace que celui pour la bonne. On peut tempérer ce résultat négatif en notant que ce qui constitue la bonne réponse à ce problème est loin d'être aussi tranché que dans les autres cas[39]. Cela est bien démontré dans une très belle analyse de Robert Nozick dans laquelle il fait monter les enjeux : le lecteur sent alors ses intuitions s'effriter et, tout rationnel qu'il soit, il préfère choisir la solution apparemment irrationnelle (car si elle s'avère être la bonne, les gains sont énormes), ce qui montre bien que les intuitions portant sur les raisons soutenant la bonne réponse ne sont pas sûres à 100% (Nozick, 1993).

Un détail peut cependant limiter la pertinence de cette étude : la provenance des arguments. Même si cela n'est pas mentionné explicitement, les participants pensent probablement qu'ils viennent de l'expérimentateur – il est donc possible qu'un phénomène d'autorité vienne biaiser les résultats, phénomène renforcé par l'aspect peu naturel, ou peu authentique des arguments (bien que ceci n'explique pas l'effet différentiel des bons et des mauvais arguments).

Afin d'éliminer ce biais et de mieux comprendre la façon dont les arguments sont produits, j'ai conduit une expérience à mi chemin entre les méthodes de la psychologie sociale et celles de la psychologie du raisonnement. Alors que les

[39] Et par ailleurs il s'agit plus d'intuitions portant sur la causalité que sur la logique.

problèmes utilisés étaient des problèmes de raisonnement classiques dans le domaine de la prise de décision, les arguments que les participants devaient évaluer n'étaient pas aussi artificiels car il s'agissait d'arguments donnés par d'autres participants. Cette méthode a permis de tester l'hypothèse, souvent faite en psychologie du raisonnement, et en particulier dans les théories à processus duel, que les participants reconnaissent, lorsqu'on leur explique, la bonne réponse aux problèmes. Mais lorsque c'est l'expérimentateur qui explique la bonne réponse, il est difficile de savoir si les participants en sont réellement convaincus, ont vraiment compris, ou s'ils ne l'acceptent que par le biais de l'autorité dont jouit l'expérimentateur. En utilisant des arguments donnés par d'autres participants ce problème ne se pose pas.

Les participants étaient tirés de populations différentes, mais il s'agit dans la quasi-totalité des cas d'étudiants, âgés d'une vingtaine d'années. Ils furent tous interrogés individuellement. Les problèmes utilisés sont les trois problèmes du 'CRT' (Cognitive Reflection Test, Frederick, 2005). Le premier est le désormais classique 'bat and ball' qui donne, dans sa version francisée, 'le bonbon et la baguette' : « Un bonbon et une baguette coûtent 1,10 € au total. La baguette coûte 1 € de plus que le bonbon. Combien coûte le bonbon ? ». Le second est le problème des nénuphars : « Dans un lac il y a une colonie de nénuphars. Chaque jour, la colonie double de taille. S'il faut 48 jours à la colonie pour recouvrir tout le lac, combien de temps faudrait-il pour que la colonie recouvre la moitié du lac ? ». Enfin, le troisième est celui des Widgets : « Il faut à 5 machines 5 minutes pour faire 5 Widgets, combien de temps faudrait-il à 100 machines pour faire 100 Widgets ? ». Le point commun à tout ces problèmes est qu'ils ont une réponse intuitive, frappante, qui vient à l'esprit de tout le monde immédiatement, mais qui est fausse (respectivement 10 centimes, 24 jours et 100 minutes). Les bonnes solutions, par contre, requièrent un minimum de réflexion, tout en étant très simples d'un point de vue mathématique, et à la portée de tous les participants (respectivement, 5 centimes, 47 jours et 5 minutes).

La méthode utilisée dépend de la 'génération' des participants. Les participants de la première génération étaient confrontés aux problèmes, un par un, devaient y réfléchir et donner une réponse. Juste après qu'ils aient donné leur réponse, ils devaient imaginer qu'ils essayaient de convaincre une autre personne que leur réponse était la bonne, leurs arguments étant enregistrés par un dictaphone. Ces arguments furent transcrits et les plus représentatifs furent très légèrement édités

pour qu'ils puissent être présentés à une deuxième génération de participants. Les participants de la deuxième génération étaient confrontés aux mêmes problèmes que ceux de la première génération, mais cette fois, après avoir donné leur réponse initiale, je leur donnais un des arguments proposé par un participant de la première génération. Ils avaient alors l'opportunité de modifier leur réponse.

Les résultats de la première génération ne sont globalement pas pertinents en eux-mêmes, ces participants (N=26) servant à produire un ensemble d'arguments qui pourraient être utilisés par la suite. Un phénomène intéressant s'est cependant produit dans une minorité de cas. Dans un dixième (4 sur 40) des cas dans lesquels les participants avaient initialement donné une mauvaise réponse, le fait de devoir donner un argument la défendant les a forcés à reconsidérer leur réponse, et ils sont parvenus à trouver la réponse correcte. Ce résultat indique que nous sommes capables d'un minimum d'autocritique vis-à-vis des arguments mêmes que nous formons, ce qui peut nous amener à reconsidérer spontanément notre solution. A partir des arguments donnés par les participants de cette première génération, quatre furent retenus pour chaque problème : deux arguments pour la bonne réponse, deux pour la mauvaise, un étant plus court et un plus élaboré pour chaque réponse. Voici par exemple les arguments retenus pour le problème du bonbon et de la baguette :

> [mauvaise réponse, court] 10 centimes, puisque la baguette coûte 1 euro et comme au total on a 1 euro 10 c'est forcément 10 centimes que coûte le bonbon.
> [mauvaise réponse, long] Le bonbon coûte 10 centimes vu qu'au total c'est 1,10, la baguette elle coûte 1 euro de plus, donc on enlève 1 euro donc ça fait 10 centimes d'euro pour le bonbon, 1 euro de plus pour la baguette, donc ça fait 1 euro 10 au total.
> [bonne réponse, court] Si on a un bonbon à 5 centimes et une baguette à 1 euro 05, on a bien un euro de plus pour la baguette, et le total fait 1 euro 10.
> [bonne réponse, long] Bon alors, je pense que le bonbon coûte 5 centimes, parce que si le bonbon coûte 5 centimes, la baguette coûte 1 euro et 5 centimes, et si on additionne les deux, on obtient 1 euro 10, et on voit que les hypothèses sont respectées, la baguette coûte 1 euro de plus que le bonbon, et on a un total de 1 euro 10.

La longueur des arguments s'avérant ne jouer aucun rôle, seul le paramètre bonne/mauvaise réponse sera conservé pour les analyses ultérieures.

Les participants de la seconde génération (N=42), après avoir donné une réponse initiale, étaient confrontés à un des quatre arguments possibles (tiré au hasard). Il était bien précisé que les arguments avaient été donnés précédemment par d'autres participants. La plausibilité était renforcée par le fait qu'ils devraient eux-mêmes, par la suite, enregistrer des arguments pour défendre leur réponse. A une seule exception près, lorsque l'argument présenté confortait le participant dans sa réponse, il ne changeait pas d'avis. Je me concentrerai donc sur les cas dans lesquels l'argument présenté au participant défendait une réponse différente de la sienne. Dans 35 cas, un participant ayant donné une bonne réponse initiale fût confronté à un argument soutenant la mauvaise réponse. Dans aucun cas un participant n'a changé d'avis suite à la présentation d'un tel argument : tous les participants conservèrent leur avis initial – à juste titre. Il y eu 40 cas en miroir, des participants ayant initialement donné la mauvaise réponse confrontés à un argument pour la bonne réponse. Dans ce cas, par contre, plus de la moitié changèrent d'avis (21). La différence entre ces deux groupes est fortement significative ($\chi^2(1) = 23{,}0$, $p<0{,}0001$). Dans ce cas, il n'y a aucune différence d'autorité entre les deux groupes : tous les arguments viennent d'autres participants. Ces résultats montrent que nombre de participants en viennent donc à accepter la bonne réponse dans les problèmes de raisonnement uniquement parce qu'ils comprennent, et acceptent, les arguments qui la soutiennent, et non par autorité.

5.2.2 **Production d'arguments**

Si la majorité des études de psychologie du raisonnement s'est focalisée sur l'évaluation d'argument, un certain nombre d'entre elles a également étudié leur production. Plus précisément, ces études ont étudié la façon dont les participants tirent des conclusions à partir de prémisses : plutôt que de leur présenter une conclusion dont ils doivent évaluer la validité, ils doivent formuler eux-mêmes une conclusion valide, ou dire qu'il n'en existe pas. Ces études ne sont cependant par pertinentes ici car il ne s'agit nullement de la façon dont les arguments sont produits au cours d'une conversation. Lorsque nous argumentons, nous partons typiquement

de la conclusion, de ce dont nous voulons convaincre notre interlocuteur, et nous cherchons des prémisses appropriées. Nous ne cherchons pas à dériver une nouvelle conclusion (et encore moins une conclusion respectant les critères stricts de la logique)[40]. La grande majorité des travaux de psychologie du raisonnement classique ne sont donc pas pertinents ici. Je me concentrerai à la place sur les études dans lesquelles les participants doivent former de réels arguments pour défendre leur position.

Les études de Perkins et Kuhn : une vision pessimiste de nos capacités d'argumentation

Dans une série d'études pionnières, Perkins et Kuhn ont demandé à des participants leur opinion sur diverses questions, et les ont poussés à défendre leur réponse. Par exemple, les participants pouvaient avoir à répondre à des réponses telles que « Would restoring the military draft significantly increase America's ability to influence world events? » (Perkins, 1985) ou « What are the causes of school failure? » (Kuhn, 1991). Les participants n'avaient que peu de temps pour réfléchir (cinq minutes dans l'expérience de Perkins par exemple), puis ils devaient donner leur conclusion et la défendre. L'expérimentateur leur demandait ensuite d'élaborer davantage, jusqu'à ce que les participants n'aient plus rien à dire.

Les conclusions de ces études sont assez cohérentes, et plutôt pessimistes vis-à-vis des capacités d'argumentation (ou de raisonnement informel comme Perkins les appelle) des participants – même des participants ayant poursuivi des études supérieures. Quelles sont les faiblesses indiquées par ces études ? La première porte sur la relative 'superficialité' des conclusions et des arguments présentés : « many reasoners could be characterized as "makes sense epistemologists." Such reasoners proceed to analyze a situation only to the point where the analysis makes superficial sense. » (Perkins, 1985, p.568). La seconde, et la principale, tient aux difficultés rencontrées dans la formation de contre-arguments, d'arguments attaquant les théories que les participants eux-mêmes ont proposées. Enfin, et cette dernière

[40] Ou alors cette nouvelle conclusion n'est pas une fin en soi, mais un élément d'un argument plus complexe – par exemple si l'on tire une conclusion peu plausible d'un énoncé de l'interlocuteur.

critique est plus spécifique à Kuhn, les participants seraient le plus souvent incapables de distinguer entre 'explications' et 'preuve'. L'exemple suivant, tiré de Brem et Rips (2000), servira à illustrer la différence entre ces deux concepts. Imaginons qu'un participant veuille défendre l'idée que « les personnes percevant des aides sociales ont du mal à s'en passer car elles ont des compétences professionnelles réduites ». Une explication peut alors être une histoire causale, un lien entre le fait de manquer de compétences professionnelles et le fait de ne pouvoir trouver un bon travail, et donc de rester dépendant des aides sociales. Une preuve, par contraste, est un élément indépendant permettant de valider la théorie. Il pourrait s'agir ici d'une comparaison statistique entre un groupe de personnes percevant des aides sociales et ayant de bonnes compétences professionnelles, et un groupe similaire mais ne disposant pas de telles compétences : si le premier groupe s'en tire mieux, il s'agit d'un soutien indépendant pour la théorie initiale.

A première vue, ces résultats sont compromettants pour la théorie argumentative : les participants semblent avoir des difficultés majeures pour former de bons arguments. Il est cependant possible de les expliquer d'une façon assez naturelle et qui, au final, apporte même un certain soutien à la théorie.

Caractère artificiel des tâches utilisées

Le premier élément sur lequel il faut insister est l'aspect très artificiel de ces tâches. Si les thèmes choisis sont des thèmes de conversation courants, et donc assez naturels, la façon de les traiter, elle, ne l'est pas du tout. Les participants ne prennent pas part à un débat : ils doivent simplement exposer leur point de vue à une personne qui les écoute et prend note, mais qui ne s'oppose jamais à eux (ni n'approuve). Il ne s'agit donc pas réellement de circonstances normales pour l'utilisation du raisonnement. Il est beaucoup plus aisé de penser à des contre-arguments lorsqu'on nous présente, effectivement, un point de vue opposé que si nous devons y penser nous-même. Par ailleurs, les thèmes, bien que communs, sont tels qu'il est très peu probable que les participants y aient porté un intérêt plus que superficiel, dépassant ce qu'ils ont pu en apprendre par les médias. On peut imaginer des centaines de thèmes similaires, et il serait absurde de s'attendre à ce qu'une personne qui n'a pas de raison particulière d'y prêter attention dispose de nombreuses informations à leur

sujet, ou y ait déjà consacré beaucoup de réflexion. Les résultats de psychologie sociale mentionnés dans la section précédente montrent qu'en évaluation, les participants n'ont recours à des stratégies sophistiquées que lorsque qu'ils sont directement motivés par le contenu des arguments. Il est plausible de tirer la même conclusion pour la production : si les participants ne sont que marginalement intéressés par le sujet sur lequel ils doivent formuler un avis et le défendre, il y a de grandes chances qu'ils aient recours à des heuristiques. Il semble donc que l'aspect 'superficiel' des arguments puisse simplement être expliqué par le peu d'intérêt que les participants y portaient (voir Stein & Bernas, 1999, pour un argument similaire).

Les limites de la capacité critique, et comment les circonvenir

La seconde critique porte sur les difficultés que rencontrent les participants à dépasser leur propre théorie. Comme le dit Kuhn, les participants ont une « tendance commune et inquiétante à assimiler toute nouvelle information aux théories existantes » (Kuhn, 1991, p.268). Par ailleurs, seul un tiers de ses participants fut capable de générer spontanément des contre-arguments. Mais étant donné les contraintes expérimentales, loin de constituer un problème pour la théorie argumentative, ce résultat rejoint au contraire ses prédictions. Dans les expériences de Kuhn et Perkins, les participants ne sont pas confrontés à un point de vue contradictoire. Bien qu'un expérimentateur soit présent, leur raisonnement est plus proche d'un raisonnement purement individuel. Dans ces circonstances, on peut s'attendre à un fort biais de confirmation (voir le chapitre suivant), biais qui n'est pas compensé par de potentiels contre-arguments qui pourraient être présentés par un interlocuteur. Il est donc tout à fait normal d'observer de tels problèmes dans ces circonstances.

Plusieurs éléments peuvent être avancés en soutien de cette interprétation. Dans une expérience de Kuhn et collègues, les participants étaient mis à la place de jurés (Kuhn, Weinstock, & Flaton, 1994). Ils étaient confrontés aux divers témoignages, à des preuves, et aux plaidoiries des avocats ; ils devaient ensuite se faire une opinion sur le verdict approprié, et la défendre. Parmi les questions qui leur étaient posées, une portait sur les verdicts alternatifs possibles. Les participants étaient donc confrontés à des conclusions possibles différant de la leur, et devaient

défendre leur propre théorie. Dans ce cas, la grande majorité des participants (84%) fournit spontanément des arguments contre les verdicts différents du leur – chiffre à comparer avec le tiers de participants qui fit de même lorsqu'aucune théorie alternative n'était proposée. Par ailleurs, plus de la majorité des participants fournit de bons contre-arguments contre d'hypothétiques arguments soutenant ces verdicts alternatifs. Utilisant les mêmes stimuli, Pennington et Hastie (1993) ont également observé de très bonnes capacités de raisonnement chez les participants lorsqu'il s'agissait de défendre leur verdict contre les alternatives. Les chercheurs ont observé que les participants étaient capable de construire des arguments non seulement convaincants, mais également valides – ou plutôt convaincants parce que valides. Par exemple, les auteurs notent que « raisonner en termes de negation (par contradiction, et par déduction négative (*modus tollens*)) est étonnamment commun » (Pennington & Hastie, 1993, p.155). Cette observation est à contraster avec le taux important d'erreur relevé lorsque des participants sont confrontés à un *modus tollens* dans un contexte non argumentatif (40% en moyenne).

Une autre étude montre que les participants sont tout à fait capables de former des contre-arguments lorsqu'ils doivent spontanément s'opposer à une conclusion. Shaw (1996) a montré à ses participants des arguments en leur demandant de penser aussi rapidement que possible à un contre-argument. Tous les participants furent capables de fournir des contre-arguments en moins de 90 secondes. Les participants étaient pourtant 'fainéants', la plupart choisissant des objections basées sur la vérité des prémisses ou de la conclusion plutôt que sur le lien entre les deux – ce qui est, dans la vie de tous les jours, parfaitement suffisant. Néanmoins, lorsque les instructions exigeaient des participants de fournir des objections basées sur le lien entre prémisses et conclusion, ils étaient plus de 80% à être encore capables d'en fournir une – toujours en moins de 90 secondes.

Enfin, on peut également mentionner une étude portant sur le raisonnement analogique (Blanchette & Dunbar, 2000). Les auteurs commencent par constater que les participants ont souvent de mauvaises performances dans les tâches de raisonnement analogique, se cantonnant à des relations superficielles entre source et objet (les deux éléments de l'analogie). Suivant le pattern général observé en raisonnement, ces mauvaises performances dans des tâches abstraites contrastent avec de bonnes performances dans la vie quotidienne : Dunbar et ses collègues ont observé que dans un contexte naturel scientifique, journalistes et politiciens utilisent

une majorité d'analogie 'profondes', faisant appel à des relations structurales entre éléments et pas simplement des relations superficielles (Blanchette & Dunbar, 2001). Afin de mettre les participants dans une situation plus naturelle, ils les ont précisément mis à la place de personnes devant défendre un programme politique face à des opposants. A travers plusieurs expériences, ils ont observé que les participants faisaient un usage très limité des ressemblances superficielles, et utilisaient au contraire une grande variété d'analogies plus profonde. Il s'agit donc d'un nouvel exemple d'amélioration de performances lié à l'usage d'un contexte argumentatif (Blanchette & Dunbar, 2000).

Explication et preuve

La troisième critique de Kuhn porte sur le manque de 'preuves' par rapport aux simples 'explications' dans les arguments des participants. Il faut souligner qu'il n'y a que très peu de chances que les participants aient eu en tête de telles preuves. Qui connaît par cœur les statistiques de récidive ou d'échec scolaire ? Les participants ont pu préférer fournir des explications car elles étaient, elles, immédiatement disponibles. D'éventuelles preuves auraient été hypothétiques : 'en prenant deux groupes, on pourrait comparer le résultat et tester ma théorie'. Ce n'est pas le genre d'argument qui est efficace dans la grande majorité des débats, dans lesquels la conclusion est décidée sur le moment, et pas après qu'une étude de grande ampleur ait été entreprise. Brem et Rips (2000) ont testé deux conséquences de cette explication.

Dans leur première expérience, ils ont demandé aux participants « to imagine the strongest supporting evidence one could provide, inventing evidence or resources for gathering evidence if necessary », et ont comparé cela aux instructions classiques utilisées par Kuhn. Si le fait que les preuves ne soient pas disponibles immédiatement freinait leur utilisation par les participants, ils devraient d'avantage y recourir dans la condition 'idéale'. C'est en effet ce qu'ils ont observé : alors que 45% des participants avançaient de réelles preuves dans la condition contrôle, ils étaient 68% à le faire dans la condition 'idéale'. Dans une autre expérience (expérience trois), ils ont modifié la disponibilité supposée des preuves possibles. Selon la condition, les participants pensaient soit que très peu d'informations étaient

disponibles sur une question donnée, ou au contraire que de nombreuses données avaient été accumulées (sans toutefois que ces informations soient directement fournies aux participants). Cette simple modification a multiplié par quatre le nombre de participants donnant des arguments basés sur des preuves (de 9% dans la condition pauvre en information, à 43,5% dans la condition riche en information). Enfin, une autre expérience (expérience deux) a montré que les participants sont également sensibles à la différence entre explication et preuve lorsqu'ils évaluent des arguments, et qu'ils préfèrent les secondes aux premières. Il semble donc bien que les gens soient sensibles à la distinction entre explications et preuves, à la fois en évaluation et en production, et que les conclusions de Kuhn étaient liées aux connaissances limitées que les participants avaient des sujets choisis, et à la difficulté de se procurer des preuves.

Analyse et effets de débats réels

J'ai mentionné plus haut qu'une des explications qu'offre la théorie argumentative pour les 'mauvaises' performances des participants dans les tâches de Perkins et Kuhn est leur aspect artificiel : il ne s'agit pas de débats, mais de l'exposition unilatérale d'un point de vue. Une conséquence de cet argument est que les performances devraient s'améliorer lorsque les gens sont observés dans de réelles situations de débat. Certaines études portent sur des textes à caractère argumentatif, ou sur les performances de 'professionnels' de l'argumentation, tels que des hommes politiques. Pour ne donner qu'un exemple, Blum-Kulka et ses collègues ont analysé des débats du Talmud et des débats entre hommes politiques Israéliens modernes (Blum-Kulka, Blondheim, & Hacohen, 2002). Dans ce cas, étant donné qu'il s'agit de spécialistes, qui plus est de spécialistes appartenant à une culture qui accorde une place importante à l'argumentation, on n'est guère surpris d'apprendre que les auteurs ont observé une « grande complexité de logique et de structure dans les arguments et les débats » (p.1569)[41].

[41] La professionnalisation et l'entraînement ne peuvent cependant que rendre compte d'une part de ces résultats. Je ne connais pas un spécialiste du raisonnement qui soit capable de résoudre les syllogismes les plus ardus avec aisance malgré des années de familiarité avec ces matériaux : donc même à niveau d'expertise équivalent, l'argumentation reste beaucoup plus naturelle que le raisonnement abstrait.

Il est plus pertinent pour mon propos de mentionner une étude portant sur des participants similaires à ceux utilisés dans les autres expériences. Resnick et ses collègues ont examiné les débats dans des groupes de trois étudiants (Resnick, Salmon, Zeitz, Wathen, & Holowchak, 1993). Les groupes avaient été ainsi constitués qu'un des membres était en désaccord avec les deux autres sur la question de l'énergie nucléaire. Au terme de leurs analyses, les auteurs concluent que :

> …we are impressed by the coherence of the reasoning displayed. Participants … appear to build complex argument and attack structure. People appear to be capable of recognizing these structures and of effectively attacking their individual components as well as the argument as a whole (Ibid, pp.362-3).

(voir Stein, Bernas, & Calicchia, 1997; Stein, Bernas, Calicchia, & Wright, 1995, pour des résultats similaires).

Enfin, on peut souligner que Kuhn elle-même note que les débats ont un effet positif sur le raisonnement. Avec ses collègues, elle a utilisé des méthodes similaires à celles mentionnées plus haut pour évaluer les capacités argumentatives des participants (question sur un sujet d'intérêt général, analyse des arguments proposés en soutien de la position défendue) (Kuhn, Shaw, & Felton, 1997). Cette fois, les capacités étaient mesurées deux fois, à un intervalle de six semaines. Dans la condition expérimentale, cet intervalle était utilisé par les participants pour discuter, par paires, du problème posé, alors que les participants de la condition contrôle ne recevaient aucun traitement particulier. Comme on peut s'y attendre, les participants de la condition contrôle ne firent aucun progrès entre les deux sessions. Par contre, le traitement expérimental permit à plus de la moitié des participants d'améliorer la qualité des arguments utilisés. Ce résultat est comparable à de nombreux autres montrant le rôle stimulant des débats pour la qualité des arguments, et la compréhension des problèmes plus généralement :

> The notion that critical thinking might be helped by peer-based teaching interventions is given support from an extensive body of research demonstrating that peer interaction is helpful in enhancing problem-solving performance and in promoting conceptual change. Peer-based learning methods are now successfully used to promote learning in a wide variety of

> domains (see the collection of papers in Foot, Howe, Anderson, Tolmie, & Warden, 1994) […] peer interaction is widely used to promote conceptual change in science learning (for example,T. Anderson, Howe, & Tolmie, 1996; Howe, 1990; Tolmie, Howe, Mackenzie, & Greer, 1993). (T. Anderson, Howe, Soden, Halliday, & Low, 2001, p.38)

(voir aussi Nussbaum & Sinatra, 2003; van Boxtel, van der Linden, & Kanselaar, 2000).

On pourrait cependant faire l'objection suivante à l'utilisation ici faite de ces études : si nous sommes naturellement bons pour raisonner dans les contextes argumentatifs, et que nous sommes spontanément portés à l'argumentation, comment se fait-il qu'un débat supplémentaire permette une telle augmentation des performances ? Les performances au contraire devraient déjà plafonner. La réponse est que l'augmentation est spécifique : elle ne s'applique qu'au domaine visé par l'entraînement. Dans les tâches citées ci-dessus, les pré- et post-tests portent sur le même sujet, ainsi que les débats servant d'entraînement. Les gens ne deviennent donc pas meilleurs pour argumenter ou raisonner généralement (ce qui est d'ailleurs montré par le fait que le niveau d'éducation n'a quasiment aucun effet, voir Perkins, 1985), ils deviennent simplement meilleurs pour trouver des arguments plus sophistiqués sur ce thème précis. Cette amélioration s'explique par le fait que leurs capacités de raisonnement ont, au moins une fois, été déclenchée naturellement en rapport avec le sujet en question, lors du débat servant d'entraînement. A l'occasion de ce débat, les capacités de raisonnement sont activées et les participants forment alors des arguments sophistiqués qu'ils utilisent ensuite lors du post-test. Ces résultats soutiennent donc l'idée que nous sommes naturellement bons pour argumenter, mais que les capacités de raisonnement nécessaires sont beaucoup plus facilement activées en contexte argumentatif.

On peut conclure cette partie en citant un résultat qui anticipe la section suivante. Une étude a utilisé les méthodes de Kuhn pour mesurer les capacités argumentatives de jurés (McCoy, Nunez, & Dammeyer, 1999). Les performances de deux groupes étaient comparées. Alors que les membres du premier groupe devaient réfléchir, seuls, sur leur verdict, les membres du second groupe débattaient sur le verdict approprié avec les autres jurés. Comme on peut s'y attendre, les membres du groupe ayant débattu étaient devenus meilleurs pour trouver des contre-arguments

visant les verdicts alternatifs. Par contre, les membres du groupe ayant eu à raisonner en solitaire étaient devenus *moins bons* : ils étaient encore plus biaisés vers leur point de vue qu'avant de devoir y réfléchir. Nous allons voir que loin d'être une exception, ce résultat reflète une caractéristique générale du raisonnement individuel.

6 Biais de confirmation

> The human understanding when it has once adopted an opinion (either as being the received opinion or as being agreeable to itself) draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects; in order that by this great and pernicious predetermination the authority of its former conclusions may remain inviolate.
>
> Francis Bacon, Novum Organum, XLCVI

La prédiction que le raisonnement *doit* être biaisé est une des plus originales de la théorie argumentative. La fonction du raisonnement n'est pas de nous aider à former de meilleures connaissances généralement, mais d'évaluer des arguments afin de déterminer les prémisses qui seront convaincantes ou les conclusions qu'il faut accepter. Lorsqu'il est utilisé en production, lorsqu'on cherche des arguments défendant une conclusion, on peut s'attendre à ce que le raisonnement soit biaisé : ce qui l'intéresse, ce sont les prémisses qui soutiennent la conclusion, pas celles qui pourraient aller dans la direction opposée. En d'autres termes, on s'attend à ce que le raisonnement montre un fort *biais de confirmation* : « seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand » (Nickerson, 1998, p.175). Bien que l'existence d'un tel biais soit invoquée pour expliquer de nombreux résultats dans des domaines divers (psychologie du raisonnement, mais aussi psychologie sociale ou prise de décision), et qu'il soit tenu responsable de certains désastres, tels que la Baie des Cochons (Janis, 1982) ou l'explosion de la navette Challenger (Kray & Galinsky, 2003), il convient d'examiner avec attention ces travaux, et ce pour deux raisons. D'une part, certains résultats qui pouvaient sembler refléter un biais de confirmation ont été réinterprétés en termes d'autres mécanismes qui, eux, ne montrent pas nécessairement de biais. D'autre part, il ne s'agit probablement pas d'un phénomène unitaire, plusieurs mécanismes peuvent être responsables d'un tel biais, et il est important de savoir lesquels sont à l'œuvre, afin de s'assurer que le raisonnement puisse bien être blâmé dans certains cas tout au moins (voir Klayman, 1995). Dans ce chapitre, nous allons revoir des travaux principalement issus de la psychologie du raisonnement qui

montrent, chacun à leur façon, que les participants ont souvent des difficultés pour remettre en cause les hypothèses qu'ils ont formées eux-mêmes, et qu'ils se contentent le plus souvent de chercher à les justifier – qu'ils font preuve, donc, d'un biais de confirmation.

6.1 Test d'hypothèse

Le 2,4,6 : biais de confirmation ***et stratégies de test d'hypothèse***

Peter Wason fut l'un des premiers à promouvoir l'idée que les gens tendent à utiliser beaucoup trop de stratégies de confirmation. Parmi les expériences qu'il a créées, une était censée refléter cette tendance avec le plus de clarté : le 2,4,6 (Wason, 1960). Dans le 2,4,6, les participants doivent trouver la règle régissant la formation de triplets de chiffres. Un premier triplet (2,4,6) est donné aux participants qui peuvent ensuite former des hypothèses sur la règle et proposer des triplets afin de la tester. L'expérimentateur indique aux participants si le triplet est conforme à la règle ou non. Les participants peuvent ensuite proposer une hypothèse sur la règle, et l'expérimentateur leur dit alors si elle est juste ou non. Les participants peuvent proposer ainsi plusieurs hypothèses avant d'arriver à la bonne (ou de se décourager). On observe généralement dans cette tâche ce qui semble être un biais de confirmation : les participants tendent à proposer des triplets qui sont conformes à leur hypothèse. Ainsi, si la première règle qui leur vient à l'esprit est 'nombres pairs se suivant', ils peuvent proposer 4,6,8, puis 10,12,14, afin de tester leur théorie. Ils ne proposent que beaucoup plus rarement des triplets qui ne correspondraient pas à la règle si leur hypothèse était bonne, tel que 2,4,8 ou 2,4,5. Or la règle à trouver est 'nombres ascendants' : les participants ont du mal à la trouver car ils se cantonnent à des hypothèses plus précises.

L'interprétation initiale en termes de biais de confirmation a cependant été remise en cause au profit de l'explication suivante (Klayman & Ha, 1987). Les participants ne chercheraient pas réellement à confirmer leur hypothèse initiale, ils utiliseraient en fait une stratégie très raisonnable de test positif. Une stratégie de test positif consiste à vérifier que quelque chose qu'on pense bon est en effet juste. Dans ce cas, il s'agit de vérifier la validité de l'hypothèse que l'on a en tête à un moment

donné en donnant des triplets s'y conformant. Cette stratégie permet de détecter les faux positifs (les cas dont on prédit qu'ils se conforment à l'hypothèse, mais qui en fait ne s'y conforment pas). Cependant, cette stratégie n'est pas efficace dans les cas de faux négatifs (les cas dont on prédit qu'ils ne se conforment pas à l'hypothèse mais qui en fait s'y conforment). Cela signifie que la stratégie de test positif sera efficace lorsque l'hypothèse envisagée est plus générale que la bonne hypothèse : une hypothèse plus générale génère en effet plus de faux positifs que de faux négatifs. Or il est assez raisonnable de s'attendre à ce que ce soit le cas le plus courant : partir d'une hypothèse générale qu'on raffine peu à peu. Le piège du 2,4,6 viendrait alors du fait que le triplet présenté amène les gens à penser à une règle très spécifique, à la fois de par la saillance de la règle et du fait que ce triplet est fournit par l'expérimentateur, ce qui renforce sa pertinence perçue (Van der Henst, 2006). Il semble donc qu'on ne puisse pas réellement parler de biais de confirmation dans ce cas, le phénomène s'expliquant par la combinaison de ces deux éléments. D'une part la grande pertinence d'une règle qui est trop générale, et d'autre part le fait qu'une fois une hypothèse en tête, les tests les plus pertinents sont des tests positifs.

On peut maintenant s'interroger sur les conditions dans lesquelles les participants sont capables de se départir de cette stratégie de test positifs qui, bien que généralement utile, s'avère être inadaptée à cette tâche. A cette fin, plusieurs auteurs ont cherché à pousser les participants à utiliser des stratégies de falsification ou d'infirmation de la règle. Ainsi, Tweney et collègues (1980) ont comparé deux conditions utilisant les instructions suivantes :

> Suppose you were shown a number triple-say, 3-3-3. You might guess that the rule governing this number triple is "three equal numbers". If you know that is the rule, you can always come up with number triples that fit that ru1e-e.g. 2-2-2; 22-22-22; and so on. . . . You could test the hypothesis by thinking of a number triple, for example, 8-8-8, and asking the experimenter if it is consistent with the correct rule. Suppose 8-8-8 is consistent with the rule. You would then have evidence supporting your hypothesis. Notice that this strategy allows you to get evidence for your hypothesis by thinking up number triples that you think would fit the rule…

> Subjects in the disconfirmation group were given instructions that were parallel to those given above, except that the example given was 5-7-9 (not 8-8-8), and the instructions continued as follows:
> Suppose 5-7-9 is consistent with the rule. You would then have evidence that your hypothesis (three equal numbers) is wrong. Notice that this strategy allows you to get evidence about whether your hypothesis is correct by testing number triples that you don't think will fit the rule… (p.114)

Avec de telles instructions, ils observèrent en effet un fort accroissement des triplets qui ne se conforment pas à la règle dans la condition infirmation (‘discomfirmation’). Etant donné la transparence des instructions sur ce point, un résultat différent eut été presque choquant. Cependant ils n’observèrent aucune amélioration des performances dans cette condition, alors même que la stratégie de vérification est censée être responsable des mauvaises performances. Pour expliquer ce résultat, il faut faire une distinction au sein même des stratégies de test positif et négatif.

Utiliser une stratégie de test positif (donner un triplet se conformant à la règle que l’on a en tête), n’est en effet pas nécessairement le signe d’une volonté de vérification : on peut parfaitement proposer un tel triplet en pensant qu’il sera refusé, prouvant ainsi que notre hypothèse était fausse. De même, une stratégie de test négatif (donner un triplet ne se conformant pas à la règle qu’on a en tête) n’est pas nécessairement falsificationniste : si le participant s’attend à ce qu’on lui réponde non, alors il s’agit d’une confirmation que sa règle est la bonne. Pour savoir si les participants utilisent bien une stratégie de falsification, il faut donc leur demander quelles sont leurs attentes. C’est ce qu’à fait Poletiek (1996) en comparant trois conditions assorties des instructions suivantes :

> Now, try to test this hypothesis with a triple [that will most likely result in rejection/confirmation of your hypothesis. Thus, try to test in such a way as to get your hypothesis about the rule rejected/supported]. (p.454)

La partie en dehors des crochets est la condition contrôle, le reste correspondant respectivement aux conditions confirmation et falsification. Les résultats montrent bien une différence entre les conditions contrôle et confirmation d’un côté, et falsification de l’autre, répliquant ainsi les résultats de Tweney et al. (1980). Cette

différence est entièrement due au fait que les participants de la condition falsification utilisèrent davantage de stratégies négatives, mais il ne s'agit cependant pas de réelles falsification : lorsqu'ils proposaient de tels triplets ne se conformant pas à l'hypothèse qu'ils avaient en tête, ces participants pensaient dans presque tous les cas (20 sur 22) que le triplet proposé allait être rejeté, confirmant ainsi leur hypothèse (ou en tout cas ne la falsifiant pas). Les instructions poussant à la falsification n'eurent aucun effet sur les vraies stratégies de falsification, à savoir le test positif avec une attente de réponse négative, et le test négatif avec une attente de réponse positive. Il semble donc que les participants aient de réelles difficultés à imaginer que leur hypothèse soit fausse.

Finalement, mentionnons une étude récente étant parvenue à augmenter substantiellement l'utilisation d'une réelle stratégie de falsification. Dans leur seconde expérience, Cowley et Byrne (2005), ont modifié le 2,4,6 en forçant la main des participants : ils leur ont fournit une hypothèse initiale, celle qui tend à être générée spontanément (nombres ascendants). Cependant, dans une condition cette hypothèse était décrite comme étant celle du participant lui-même, alors que dans l'autre elle était censée être celle d'un autre participant. Les participants tentèrent beaucoup plus fréquemment de falsifier l'hypothèse lorsqu'elle était présentée comme venant d'autrui que comme étant la leur[42] et l'abandonnèrent également beaucoup plus aisément[43]. Cela montre que si les gens ont bien du mal à concevoir que leurs hypothèses puissent être fausses, ils n'ont pas de telles difficultés (en tout cas beaucoup moins) lorsqu'une hypothèse identique est proposée par quelqu'un d'autre. Il me semble difficile d'expliquer ces résultats par l'utilisation de stratégies qui ne soient pas biaisées d'une façon ou d'une autre. Il n'est cependant pas évident de déterminer à quel stade et dans quelles conditions intervient ce biais. Est-ce que les gens ne cherchent qu'à confirmer leurs propres hypothèses mais adoptent une stratégie plus normative lorsqu'il s'agit de l'hypothèse de quelqu'un d'autre ? Au contraire, peut-être leur stratégie individuelle est-elle optimale, et sont-ils par trop méfiants vis-à-vis des hypothèses alternatives ? Ou encore sont-ils biaisés dans les deux sens, étant trop protecteurs vis-à-vis de leurs hypothèses en même tant que trop

[42] Ils furent plus de quatre fois plus nombreux à le faire, bien que cette différence n'ait pas été significative, probablement à cause des effectifs limités : N=16 par condition.

[43] Une différence cette fois hautement significative.

critiques vis-à-vis des autres ? C'est cette dernière solution que prédit la théorie argumentative, même si, étant donné la difficulté de déterminer ce qui représente une stratégie optimale, il reste très ardu de trancher entre ces hypothèses.

Autres exemples de test positif d'hypothèse

Dans d'autres domaines que la psychologie du raisonnement des auteurs ont relevé cette tendance à recourir à des tests positifs, ou plutôt, à des vérifications et non des falsifications. Avant de passer à des résultats issus de la psychologie sociale, on peut mentionner une étude qui tente de simuler le test d'hypothèse dans le travail scientifique (Mynatt, Doherty, & Tweney, 1977). Dans cette étude les participants devaient former des hypothèses sur la façon dont différents objets (des formes sur un écran) affectaient le trajet d'une particule (qu'ils voyaient se déplacer sur ce même écran). Après une série d'observations dans lesquelles la trajectoire de la particule était influencée par les différents objets présents, ils devaient former une hypothèse. Ensuite, on leur présentait par paires des écrans sur lesquels ils seraient susceptibles de tester leurs hypothèses. Ils ne devaient retenir qu'un écran par paire. Or les paires étaient construites de telle façon que dans la moitié des cas un des écrans correspondait à une stratégie de confirmation (il permettrait potentiellement de confirmer l'hypothèse des participants) et l'autre non. Voici un exemple. La première partie de l'expérience était construite pour mettre un grand nombre de participants sur la piste d'une hypothèse selon laquelle les triangles auraient une influence prépondérante sur la trajectoire des particules. Une paire d'écrans pouvait ensuite être composée d'un triangle d'un côté, et d'un carré de l'autre. La stratégie de confirmation consiste alors à choisir l'écran avec le triangle : si la trajectoire de la particule sur cet écran est bien influencée par l'objet, cela confirmera l'hypothèse – alors que l'écran avec le carré ne pourrait que l'infirmer si la trajectoire était là aussi influencée par l'objet.

Le problème de cette expérience est qu'on peut en expliquer les résultats par des effets de bas niveau de pertinence (similaires au biais d'appariement que nous examinerons dans la partie consacrée à la tâche de sélection de Wason). Les participants pourraient ne pas réellement chercher à confirmer leur hypothèse, ou éviter de l'infirmer, mais se contenter de choisir les écrans qui sont les plus

pertinents étant donné leur hypothèse – ceux qui comportent des éléments mentionnés dans leur hypothèse. Il n'est pas nécessaire d'invoquer des stratégies de raisonnement pour rendre compte de ces résultats, ce qui ne veut pas dire qu'ils ne reflètent pas des mécanismes psychologiques intéressants, mais que si biais il y a, il ne se situe pas nécessairement au niveau qui nous intéresse ici. De plus, et comme les auteurs l'indiquent, on peut contester l'aspect non normatif de ces choix. Imaginons que vous vouliez tester l'hypothèse selon laquelle un certain produit chimique facilite l'oxydation d'un métal donné. Il paraît normal de commencer par expérimenter avec ce métal et pas avec d'autres. De plus, dans l'expérience de Mynatt et al., les participants réagissent de façon adéquate lorsque les résultats de ce second test sont connus : s'ils falsifient leur hypothèse, les participants sont capables (10 sur 11) de la rejeter. Il n'est donc pas clair que ces résultats représentent une stratégie de vérification abusive.

Une série d'expériences dont les conclusions sont similaires bien que le domaine soit assez différent a été menée à bien par Snyder et ses collègues. Dans la première de ces expériences (M. Snyder & Swann, 1978), les participants devaient choisir des questions qu'ils pensaient poser lors d'une interview censée déterminer le caractère d'une personne. Dans une condition les participants devaient chercher à déterminer, en choisissant les questions appropriées, si la personne était extravertie. Pour les aider, il leur était donné le texte suivant afin de les aider :

> Extraverts are typically outgoing, sociable, energetic, confident, talkative, and enthusiastic. Generally confident and relaxed in social situations, this type of person rarely has trouble making conversation with others. This type of person makes friends quickly and easily and is usually able to make a favorable impression on others. This type of person is usually seen by others as characteristically warm and friendly.

Dans l'autre condition, ils devaient déterminer si la personne était introvertie, un texte similaire décrivant ce type de personnalité leur étant fournit à la place. Ils devaient ensuite choisir des questions parmi une liste qui était construite de telle sorte que certaines questions présupposaient une personnalité extravertie ('What would you do if you wanted to liven things up at a party?') et d'autres une

personnalité introvertie (‘In what situations do you wish you could be more outgoing?’). Le résultat principal fut que les participants devant déterminer si une personne était extravertie avait tendance à choisir le premier type de questions, alors que ceux qui devaient déterminer si une personne était introvertie choisissaient le second. Pour les auteurs, il s’agit d’un biais vers des stratégies de confirmation. Si on reprend cependant les distinctions introduites dans la discussion du 2,4,6, on réalise qu’il ne s’agit en fait que d’un biais vers des tests positifs. Il faudrait donc un élément supplémentaire : les attentes des participants. Si les participants s’attendent à ce que les personnes interviewées répondent positivement aux questions, il s’agit alors d’un vrai biais vers des stratégies de vérification et non de falsification. S’il n’y a pas de preuve définitive de ceci, la façon dont les questions sont posées est très suggestive de l’attente d’une réponse positive : presque tout le monde connaîtra au moins un moyen d’égayer une fête, et il serait très surprenant que les participants choisissant cette question s’attendent à ce que la personne leur dise qu’elle n’en connaît aucun (ce qui correspondrait alors à une stratégie de falsification)[44].

Même en admettant qu’il s’agisse bien de stratégies de vérification, on peut douter que le raisonnement en soit responsable. A nouveau, il me semble plus probable que des phénomènes de pertinence cognitive (Sperber & Wilson, 1995) n’impliquant pas nécessairement le raisonnement soient en jeu. Les participants viennent en effet de lire un paragraphe décrivant assez précisément les comportements d’un type de personne. On peut dès lors penser que les questions qui portent sur ces comportements (et pas sur des comportements non mentionnées) seront jugées plus pertinentes pour les sujets sans que le raisonnement ne doive introduire de biais supplémentaires.

On retrouve une situation très similaire dans une étude suivante (M. Snyder & Cantor, 1979). Cette fois, les participants commençaient par lire la longue description du comportement d’une personne. Deux jours plus tard, ils avaient pour tâche de déterminer si cette personne serait qualifiée pour exercer un emploi donné. Dans une condition cet emploi correspondait à un stéréotype introverti (bibliothécaire) et l’autre à un stéréotype extraverti (travailler dans une agence

[44] Et j’ai une forte intuition que quelqu’un qui poserait la question en s’attendant à une réponse négative pense en fait que la personne est très introvertie, auquel cas il s’agirait encore d’une stratégie de vérification.

immobilière). De même que dans l'étude précédente, un descriptif assez précis du type de personnalité convenant pour le poste était fourni aux participants, qui devaient ensuite rapporter tous les éléments de la description du comportement de la personne dont ils se souvenaient et qui étaient pertinents pour évaluer la qualification de la personne pour le poste. Le résultat attendu fut obtenu : les participants qui devaient évaluer si la personne était qualifiée pour un poste requérant une personnalité extravertie se souvinrent de plus d'éléments reflétant l'extraversion, et vice versa. Cependant, on peut également douter qu'il s'agisse d'autre chose que d'un résultat dû à des effets de pertinence : certains éléments de l'histoire sont rendus plus pertinents car ils sont proches de la façon dont est décrite la personnalité recherchée (par exemple 'Jane engaged in animated conversation in the doctor's office' et 'talkative').

En conclusion de cette section sur le test d'hypothèse, on peut dire que la plupart des résultats régulièrement utilisés dans la littérature pour montrer l'existence d'un biais de confirmation peuvent en fait être interprétés d'une façon qui semble disculper le raisonnement. Il s'agirait plutôt de phénomènes de pertinence qui font que des éléments proches de ce qui est à l'esprit des participants à un moment donné sont sélectionnés. Il se trouve par ailleurs qu'on peut considérer que ces éléments servent une stratégie de confirmation, mais si ça peut être là leur effet (comme ça l'est dans le 2,4,6 ou les autres expériences passées en revue ici), ce n'est pas forcément leur objectif. Deux éléments font que ces explications ne sont pas parfaitement satisfaisantes cependant. D'une part, elles n'expliquent pas certains des résultats obtenus, comme la difficulté, même lorsque les instructions sont explicites, d'utiliser des stratégies de falsification, alors même que les participants y ont recours assez spontanément lorsque l'hypothèse à examiner n'est pas la leur – résultats qui par ailleurs sont conformes aux prédictions que l'on pourrait faire sur la base de la théorie argumentative. D'autre part, même si ces stratégies n'ont pas un réel objectif de confirmation, c'est bien là leur effet, et un tel effet ne sera guère adaptatif dans bon nombre de situations. Si le principe de pertinence est destiné à améliorer notre fonctionnement cognitif, il est étrange qu'il ait de telles conséquences. Je reviendrai plus longuement sur ce problème après avoir revu d'autres résultats qui permettent de faire une distinction plus claire entre biais causés par des phénomènes de pertinence et biais causés par le raisonnement lui-même.

6.2 Tâche de sélection de Wason

Revenant à des expériences de psychologie du raisonnement, les résultats d'une autre tâche introduite par Peter Wason furent d'abord interprétés comme un cas de biais de confirmation. La tâche de sélection de Wason a déjà été introduite (chapitre 3). La première version de la tâche abstraite indiquait que les participants ne choisissaient pas les cartes qui falsifieraient la règle, mais au contraire celles qui la vérifiaient. Cette interprétation dut cependant rapidement être révisée lorsque les résultats de nouvelles variantes furent découverts. Lorsque la règle comporte une négation dans le conséquent, les participants ont de très bonnes performances (voir Evans, 1998). Non qu'ils utilisent alors une réelle stratégie de falsification, mais le simple fait de sélectionner les cartes mentionnées dans la règle leur permet de donner la réponse correcte. Ce résultat, et d'autres, a donné lieu à une explication en termes de biais d'appariement : les participants se contenteraient de sélectionner les cartes qui sont mentionnées dans la règle (même dans la portée d'une négation) (Evans, 1998). Cette explication a été rendue caduque par une interprétation plus générale de la tâche en termes pertinentistes (Sperber et al., 1995). Selon cette interprétation, le contexte dans lequel la règle est énoncée, ainsi que le contenu de la règle rendent certaines solutions pertinentes, et ce sont ces solutions que les sujets choisissent : il s'agit parfois des bonnes, parfois des mauvaises, mais le raisonnement ne semble pas être vraiment impliqué, au moins dans le stade d'orientation initial vers les cartes à sélectionner.

Voici un exemple de problème utilisé par Sperber et al. :

> Until recently, it was obvious that a woman who has children has had sex. With artificial insemination, it is now possible for a virgin to have children. The leader of the Haré Mantra (a very secret religious, Californian sect) has been accused of having had some of his sect's virgin girls artificially inseminated.
>
> His goal, it is claimed, is to create an elite of "Virgin-Mothers" alleged to play a particular role in his religion. The head of the Haré Mantra makes a joke out of these allegations. He claims that the women of his sect are,

without exception, like any other women: if a woman has a child, she has had sex.

Imagine that you are a journalist and that you are preparing an article on the Haré Mantra. You learn that a gynecological survey has been carried out among the Haré Mantra women. Some of them might be "Virgin Mothers". You go and visit the doctor who carried out the gynecological survey. Unfortunately, the doctor pleads professional secrecy and refuses to tell you what he discovered.

You realise that, before you on the doctor's desk, there are four individual information cards about Hare Mantra women examined in the gynecological survey. However, these four cards are partially concealed by other papers (as shown below). Of two cards, one can only see the top where you can read whether the woman has children or not. Of the other two cards, you can only see the bottom where you see whether the woman has had sex or not. You are determined to take advantage of a moment in which the doctor turns his back to uncover the papers and to learn more.

Indicate (by circling them) those cards that you should uncover in order to find out whether what the leader of the Hare Mantra says ("if a woman has a child, she has had sex") is true, as far as these four women are concerned, indicate only those cards that it would be absolutely necessary to uncover.

Les quatre cartes mentionnent respectivement ‘enfant : non’, ‘enfant : oui’, ‘sexe : non’ et ‘sexe : oui’. Les auteurs font l’hypothèse que dans ce cas la possibilité de ‘mère-vierge’ sera rendue très pertinente, et donc que les cartes correspondantes (‘enfant : oui’ et ‘sexe : non’), qui correspondent en effet à la bonne réponse, seront choisies par de nombreux participants. Ce fut le cas : une grande majorité de participants (78%) donna cette réponse dans cette condition (à comparer aux 10% de bonnes réponses obtenus habituellement dans la tâche standard). Dans la mesure où cette théorie rend compte des résultats par des effets majoritairement d’ordre pragmatiques, on pourrait penser que les résultats n’ont que peu à voir avec le raisonnement. On peut cependant se demander si une partie de la pertinence n’est pas déterminée par des facteurs argumentatifs. Dans l’exemple ci-dessus, le contexte de l’histoire rend pertinentes les ‘mères-vierges’, mais celles-ci sont également

pertinentes *en tant qu'argument* : le fait de trouver de telles 'mères-vierges' permet de réfuter ce que dit le responsable de la secte, une tâche que les participants ont probablement a cœur de mener à bien. On retrouve la même possibilité d'interprétation en termes de pertinence argumentative dans d'autres problèmes utilisés : il est difficile de déterminer dans quelle mesure une solution donnée est rendue pertinente pour des raisons argumentatives ou uniquement pour d'autres raisons pragmatiques et cognitives (notons que les explications pragmatiques et cognitives jouent un rôle dans tous les cas, la question est uniquement de savoir si des considérations de pertinence *en tant qu'argument* jouent également un rôle).

Bien qu'il ne soit pas possible de trancher dans le cas de ces expériences, une autre étude menée à bien avec cette même tâche, mais dans le cadre des recherches sur le raisonnement motivé, indique que la pertinence de certaines réponses en tant qu'argument peut jouer un rôle (Dawson, Gilovich, & Regan, 2002). Dans la première expérience, les participants commençaient par remplir un questionnaire de stabilité émotionnelle ayant la propriété d'avoir des résultats très tranchés (les gens sont facilement catégorisés comme émotionnellement stables ou instables). Ensuite, on leur expliquait les résultats de recherches liant cette mesure à la durée de vie. Dans une condition, une forte stabilité émotionnelle tendait à mener à une mort précoce, alors que dans une autre il s'agissait de l'instabilité émotionnelle. Ensuite, les participants des deux conditions étaient confrontés à quatre cartes représentant les résultats d'une nouvelle étude sur le sujet. Chaque carte correspondait au résultat d'un sujet de l'étude. Une carte mentionnait une forte stabilité émotionnelle, une autre une instabilité, une troisième une mort précoce et la dernière une mort tardive. Les participants devaient vérifier si ces sujets se conformaient bien à la règle qui leur avait été donnée (que la stabilité [instabilité] émotionnelle amenait à une mort précoce). Dans ce cas, les participants émotionnellement stables et à qui on avait dit que cette stabilité pouvait entraîner une mort précoce devraient être motivés pour prouver que la règle est fausse : ils devraient alors retourner les bonnes cartes. On devrait observer le même phénomène pour les participants instables qui pensent que cela peut être annonciateur d'une mort précoce. A l'inverse, les deux autres groupes (ceux dont la règle indique qu'ils devraient vivre longtemps étant donné leur stabilité ou instabilité émotionnelle) ne devraient pas du tout être motivés pour réfuter la règle. Les résultats furent conformes aux prédictions : alors que 10% des participants

non motivés trouvèrent la bonne solution, près de la moitié de ceux qui étaient motivés la trouvèrent.

Ce résultat est intéressant dans ce contexte car il me semble difficile d'établir des différences de pertinence intrinsèques entre les conditions. En effet, au sein de chaque condition les mécanismes cognitifs et pragmatiques de pertinence devraient être conservés : les instructions, le contexte, les cartes sont les mêmes. La seule différence tient à la motivation des participants à montrer que la règle est fausse. Dans ce cas, les participants sont capables de sélectionner les bonnes cartes car il s'agit d'éléments (d'arguments) pouvant contribuer à montrer que la règle est fausse, et non à cause de phénomènes de pertinence sans lien avec le raisonnement.

La seconde expérience menée à bien par ces auteurs confirme ce résultat. Les participants devaient tout d'abord donner un stéréotype concernant un groupe ethnique auquel ils appartiennent (par exemple : « les Africains-Américains sont doués en musique »). Dans une condition ce stéréotype était positif, il était négatif dans une autre. Ensuite, les participants devaient tester la validité de cette règle pour quatre individus, qui étaient alors représentés par l'expérimentateur de façon à correspondre à une tâche de sélection classique (un individu appartenant au groupe et un non, un individu se conformant au stéréotype et un non). Dans le cadre du raisonnement motivé, on s'attend au résultat suivant : seuls les participants appartenant à un groupe qualifié par un stéréotype négatif devraient trouver les bonnes cartes, celles qui permettent de falsifier la règle. Les autres, qu'il s'agisse de ceux appartenant à un groupe qualifié d'un stéréotype positif, ou ceux n'appartenant pas au groupe, ne devraient pas être autant motivés et devraient donc tomber dans les travers normaux de la tâche. C'est bien ce qui fut observé : alors que les participants non motivés stagnaient entre 14 et 20% de bonnes réponses, les participants motivés atteignirent 52%. Cette deuxième expérience confirme l'importance que peut prendre la pertinence des réponses en tant qu'arguments : lorsque les participants sont motivés pour montrer que la règle est fausse, les éléments qui permettent de l'infirmer deviennent pertinents alors qu'ils ne le sont pas pour d'autres raisons.

Cet exemple fournit une bonne illustration du fonctionnement du raisonnement. Pour la théorie argumentative, le raisonnement nous permet soit de rechercher directement, soit de filtrer des propositions de façon à retenir celles qui sont pertinentes en tant qu'argument. Donc, lorsqu'un participant donne une réponse

qui n'est pertinente que comme argument et d'aucune autre façon, cela montre que le raisonnement a joué son rôle.

Si on considère à la fois ces résultats et les résultats précédents portant sur l'importance plus générale des phénomènes de pertinence dans la tâche de sélection, on peut en conclure que des processus n'ayant rien à voir avec le raisonnement sont en bonne partie (parfois totalement) responsable du choix initialement fait par les participants, mais également que dans les bonnes circonstances, lorsque les participants sont adéquatement motivés, le raisonnement peut leur permettre de donner la bonne réponse. Cela ne signifie cependant pas que le raisonnement n'est pas du tout utilisé par les participants qui ne sont pas motivés pour montrer que la règle est fausse et qui donnent la mauvaise réponse. Cela est immédiatement visible chez la plupart des participants qui passent souvent longtemps à réfléchir sur la tâche. Dans la mesure où les mécanismes de pertinence orientent l'attention des participants vers certaines cartes de façon quasi immédiate, on peut se demander ce qu'ils font pendant ce temps. Il existe des méthodes permettant à la fois de s'assurer que les participants commencent bien par considérer les cartes jugées les plus pertinentes et d'en savoir plus sur ce qu'ils font ensuite. On peut ainsi leur demander d'énoncer leurs pensées à haute voie (Lucas & Ball, 2005), de placer un curseur sur la carte à laquelle ils sont en train de réfléchir (Evans, 1996), ou encore suivre leur regard en supposant qu'il se porte sur la carte à laquelle ils sont en train de penser (Roberts & Newton, 2002).

Il n'est pas nécessaire de détailler ces études car leurs résultats convergent fortement. Tout d'abord, les participants commencent bien par se focaliser sur les cartes les plus pertinentes. Mais le plus intéressant est leur comportement suivant : ils passent plus de temps à réfléchir aux options qu'ils vont effectivement choisir, celles vers lesquelles ils se sont orientés dès le début. Les auteurs interprètent cela comme des tentatives de rationalisation : les participants chercheraient à justifier leurs réponses. Il est dur de ne pas voir ici une forme de biais de confirmation : dans la mesure où les participants passent tout de même un peu de temps à réfléchir sur chaque carte (même s'ils en passent plus sur certaines), pourquoi ne parviennent-ils pas à trouver la bonne réponse ? Prenons par exemple le cas d'un participant donnant la réponse typique de retourner les cartes *p* et *q* (pour une règle *si* p *alors* q, et alors que la bonne réponse est *p* et *non-q*). Il aura tendance à justifier le choix de *q* en disant « si on retourne la carte et qu'il y a un *p*, alors la règle est vraie ». Cependant,

il est incapable de mettre en place un raisonnement semblable pour la carte *non-q* : « si on retourne la carte et qu'il y a un *p*, alors la règle est fausse ». Dans la mesure où ces deux raisonnements sont de complexité comparable, il est difficile d'expliquer le fait que les participants accèdent au premier et non au second à cause de limitations cognitives[45]. Il semble bien, au contraire, qu'il faille recourir à un biais : les participants cherchant surtout des arguments soutenant leur choix initial, ils ont beaucoup plus de chances de trouver le premier que le second.

6.3 Syllogismes

Introduction

Les syllogismes ont une longue histoire dans le domaine de la psychologie, et ils sont encore un des objets d'étude centraux de la psychologie du raisonnement. Ils furent parmi les premiers à être étudiés, par Störring au début du XX^{ème} siècle (Störring, 1908). Les syllogismes classiques ont deux prémisses dont les deux termes extrêmes sont combinés dans une conclusion à l'aide du terme moyen, comme dans l'exemple suivant :

> Tous les athéniens sont des grecs
> Tous les grecs sont mortels
> Donc tous les athéniens sont mortels

Depuis Störring, le raisonnement syllogistique est devenu un des terrains sur lesquels s'affrontent différentes théories du raisonnement, et la littérature sur le sujet est maintenant conséquente, sans pour autant qu'un consensus se dégage (loin de là). L'objectif de cette partie n'est pas de proposer une théorie précise du raisonnement syllogistique : cela impliquerait par exemple de faire des hypothèses très précises sur la sémantique et la pragmatique des quantificateurs utilisés (tous, certains, aucun). Je

[45] Bien qu'une réponse classique pourrait être de dire que le *modus tollens* est trop difficile pour les participants, on sait que ceux-ci peuvent le comprendre (V.A. Thompson, Evans et al., 2005) et le produire (Pennington & Hastie, 1993) sans aucune difficulté dans les bons contextes. S'ils ne le font pas dans ce cas, ce n'est donc pas à cause de difficultés liées à la complexité du *tollens* en elle-même.

me contenterai donc de faire des prédictions assez générales (dans l'ensemble, certaines seront tout de même plus spécifiques) tirées de la théorie argumentative, et de tirer de la littérature des éléments pertinents. Ceci permettra de montrer de nouveaux exemples de phénomènes qu'on peut assimiler au biais de confirmation. Mais avant de faire ces prédictions, il est nécessaire d'exposer un peu plus avant une des théories du raisonnement syllogistique car c'est dans ses termes que les prédictions seront exposées (cela ne reflétant pas une adhésion à la théorie, mais il serait inutile de développer un nouveau vocabulaire pour décrire des phénomènes communs).

La théorie des modèles mentaux (voir Johnson-Laird, 1999, 2001 pour des revues récentes) postule que nous formons des modèles des différentes situations décrites par le langage (ou perçues), modèles sur lesquels nous pouvons ensuite effectuer certaines opérations en vue de parvenir à une conclusion valide. Ces modèles sont composés d'exemplaires (tokens) qui sont dans une relation de correspondance avec des éléments de la situation. L'organisation de ces exemplaires dans notre espace mental représente les relations entre les différents éléments de la situation. Afin d'éclaircir ces concepts, voici une illustration du modèle initial évoqué par la phrase « tous les M sont des P » :

[m] p
[m] p

Les « m » et les « p » représentent ici des exemplaires appartenant aux classes M et P. Les crochets (une « note mentale », pour reprendre la terminologie de Johnson-Laird) entourant les deux « m » symbolisent le fait que ces exemplaires sont représentés exhaustivement : ici cela signifie qu'un « m » ne peut apparaître nulle par ailleurs. Notons qu'il ne s'agit que d'un modèle initial, modèle qui pourra ensuite être sujet à révision.

Après avoir évoqué la manière dont des situations sont représentées dans la théorie des modèles mentaux, tournons-nous vers son application aux syllogismes. Le raisonnement syllogistique fut le premier domaine couvert par cette théorie (Johnson-Laird & Steedman, 1978). Depuis, la théorie n'a pas cessé de changer au travers des multiples formulations qui en ont été données afin de mieux accommoder les résultats empiriques (par exemple Bara, Bucciarelli, & Johnson-Laird, 1995;

Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991; Johnson-Laird & Byrne, 1996). Je m'appuierai ici sur la dernière formulation proposée, celle de Bucciarelli et Johnson-Laird (1999).

Lorsqu'un participant est confronté à une paire de prémisse, la première étape consiste à créer un modèle représentant chacune des prémisses, de la manière décrite ci-dessus. Il doit ensuite combiner les modèles obtenus afin de parvenir à une représentation unique de la situation. Pour cela, les exemplaires représentant le terme moyen sont mis en commun entre les deux modèles, comme dans l'exemple suivant :

Tous les M sont des P	Tous les S sont des M
[m] p	[s] m
[m] p	[s] m

modèle intégré :

[s] [m] p
[s] [m] p

A nouveau, ce modèle n'est qu'un modèle initial. Il permet de formuler une ou des conclusions temporaires (ici, les conclusions possibles sont « tous les S sont des P » et « tous les P sont des S »), que le participant devra ensuite chercher à infirmer dans la troisième phase du processus, la construction de contre-exemples.

Le participant a donc une conclusion qu'il va essayer de réfuter. Pour cela, il dispose selon Bucciarelli et Johnson-Laird (1999, p.254) de trois méthodes : il peut déplacer certains des exemplaires à l'intérieur du modèle, ajouter des exemplaires au modèle et joindre ensemble deux entités. Ces techniques permettent de rechercher de manière exhaustive l'ensemble des contre-exemples possibles, de telle manière que si elles sont accomplies parfaitement, la réponse correcte est toujours fournie. Cependant, il peut être plus ou moins difficile de trouver ces contre-exemples, et la difficulté de la tâche est relative au nombre de modèles nécessaires à la recherche exhaustive des contre-exemples. Reprenons l'exemple ci-dessus. La seule méthode de réfutation possible consiste à ajouter un exemplaire (les exemples sont tirés de Bucciarelli et Johnson-Laird, 1999) :

[s] [m] p

[s] [m] p
p

Il est alors possible d'éliminer la conclusion « tous les P sont des S », pour ne conserver que la conclusion valide « tous les S sont des P ». Dans ce cas, il a suffit d'un modèle pour parvenir à cette conclusion, le problème est donc jugé comme étant facile. Par contre, d'autres syllogismes requièrent la construction de plusieurs modèles. Voici par exemple le modèle initial du syllogisme « certains P sont des M / aucun M n'est un S » :

p [m] –s
p
[m] –s
[s]
[s]

Il est possible de tirer les deux conclusions suivantes : « aucun P n'est un S » et « aucun S n'est un P ». Un premier modèle visant à réfuter ces conclusion peut être crée en déplaçant un exemplaire :

p [m] –s
p [s]
[m] –s
[s]

Ce modèle réfute les deux conclusions, mais permet encore leurs implications : « certains S ne sont pas des P » et « certains P ne sont pas des S ». Un dernier modèle est alors construit, cette fois en ajoutant un exemplaire au modèle :

p [m] –s
p [s]
[m] –s
p [s]

La seule conclusion qui persiste est alors « certains P ne sont pas des S ».

Cette présentation permet de faire deux prédictions : d'une part, les syllogismes nécessitant la construction d'un seul modèle seront plus faciles à résoudre que ceux qui en requièrent plusieurs ; d'autre part, les erreurs faites par les participants devraient refléter les modèles initiaux qu'ils ont représentés. Ces prédictions ont été vérifiées avec un succès variable dans les multiples expériences menées par Johnson-Laird et ses collaborateurs (voir les nombreuses références citées plus haut), le résultat le plus robuste reposant sur la distinction entre syllogismes à un et à plusieurs modèles, les premiers étant généralement plus faciles que les seconds.

Voici donc la théorie des modèles mentaux dans ses grandes lignes, telle qu'elle est appliquée au raisonnement syllogistique. Quelles sont maintenant les prédictions que pourrait faire la théorie argumentative ? Tout d'abord, la prédiction la plus générale concerne le degré d'implication du raisonnement dans l'évaluation de la conclusion (pour ce qui est des tâches d'évaluation – celles dans lesquelles une conclusion est donnée – à contraster aux tâches de production dans lesquelles les participants doivent former eux même la conclusion). Etant donné qu'il ne s'agit absolument pas d'un contexte argumentatif, que personne n'essaie de convaincre les participants de la conclusion (qui de toute façon ressemble généralement à « tous les A sont des C » ou « tous les horticulteurs sont des footballeurs »), le raisonnement ne devrait être que très faiblement activé dans cette première phase d'évaluation. Dans les tâches de production, on peut même penser que le raisonnement ne sera pas activé du tout lors de la phase initiale : étant donné que l'expérimentateur n'a aucun dessein de conviction dont le participant pourrait se servir pour inférer une conclusion implicite, le raisonnement n'a pas de raison d'être utilisé.

De même que pour la tâche de sélection de Wason, les participants devraient avoir une intuition initiale sur la réponse à donner, intuition due à différents facteurs mais qui peut s'exprimer, par exemple, en termes de modèles mentaux : il s'agit du premier modèle créé en intégrant les prémisses. Nous verrons ensuite que d'autres facteurs peuvent jouer un rôle à ce niveau (tels que la crédibilité de la conclusion). Cependant, et à nouveau de même que dans la tâche de sélection de Wason, on peut prédire que les participants souhaiteront ensuite chercher des justifications, des

arguments pour leur réponse[46]. Dans ce cas, bien qu'ils utilisent le raisonnement, les résultats devraient être très différents de ceux prédits par la théorie des modèles mentaux : les participants ne devraient en effet pas chercher à falsifier la conclusion formée initialement, mais au contraire à la justifier.

Syllogismes : manque d'engagement

Il n'est pas aisé de mesurer l'engagement général des participants dans la tâche. Deux résultats, un très général et un qui peut sembler presque anecdotique, indiquent néanmoins que l'engagement est minimal. Tout d'abord, bien qu'il y ait un effet global de la validité (les conclusions des syllogismes logiquement valides tendent à être plus acceptées que celles des syllogismes non valides), les performances sont loin d'être parfaites. Ainsi, certains syllogismes sont résolus par moins de 10% des participants (voir Geurts, 2003, pour une revue des résultats), cela malgré le fait que les problèmes ne sont pas ardus d'un point de vue computationnel. Les participants sont des étudiants qui ont tous eu à résoudre des problèmes beaucoup plus complexes au cours de leur scolarité. Il ne s'agit pourtant pas d'un manque de motivation pour ce qui est de réussir la tâche : dans les expériences que j'ai conduites, les participants sont anxieux de savoir s'ils ont bien réussi, pensant souvent qu'il s'agit d'une forme de test d'intelligence. Il semble plutôt qu'il s'agisse d'un manque d'implication naturelle du raisonnement, pour ce qui est de la phase d'évaluation initiale de la conclusion tout au moins.

Un autre élément montrant que le raisonnement ne fournit pas aux participants une base solide pour répondre est l'influence de facteurs extra-logiques. Le cas de la crédibilité des conclusions sera examiné plus tard, car il n'est au contraire pas surprenant qu'il soit pris en compte (dans le cadre de la théorie argumentative au moins). Par contre, il est plus surprenant que le taux de base d'acceptation des conclusions joue un rôle. Dans certaines de leurs expériences, Klauer, Musch et Naumer (2000, voir expérience 7 en particulier) ont utilisé des syllogismes dont les conclusions sont presque toujours acceptées (ou rejetées) par les participants afin de faire varier le taux de base d'acceptation et de rejet des

[46] En admettant qu'ils ne soient pas trop ennuyés par la tâche après leur 10ème syllogisme, auquel cas ils pourraient très bien se contenter de leur intuition initiale.

conclusions. Il se trouve que ce taux de base a un effet assez important sur les réponses à d'autres syllogismes, plus ardus : certains syllogismes voient ainsi leur taux d'acceptation augmenter de 15 ou 20% entre une condition avec un faible taux d'acceptation de base et une condition avec un taux élevé. Il est possible que cette manipulation ait eue un effet au niveau intuitif (rendant l'acceptation plus accessible), ou au un niveau plus réflexif, rendant la justification « il y a l'air d'avoir beaucoup de réponses positives dans cet exercice, je vais répondre oui à celui-ci aussi » plus accessible, et ainsi orientant vers une acceptation dans les cas ou les intuitions sont faibles. Quoi qu'il en soit, le fait qu'un facteur aussi non pertinent soit pris en compte montre bien que les capacités naturelles de raisonnement des participants ne sont pas fortement engagées par ces tâches.

Syllogismes : absence de falsification

Si, comme nous venons de le voir, le raisonnement ne joue qu'un rôle mineur dans une première phase d'examen ou de construction de la conclusion, il peut ensuite être utilisé pour chercher à la justifier. A l'inverse, la théorie des modèles mentaux prédit que si le raisonnement est utilisé à ce moment, ce devrait être pour essayer de falsifier la conclusion initialement retenue, en construisant des modèles alternatifs. Afin de tester l'éventuelle construction de ces conclusions alternatives Newstead et ses collègues (1999) ont conduit une expérience identique aux expériences classiques sur les syllogismes, à un détail près. Après que les participants ont donné la conclusion qu'ils pensaient être valide (ou indiqué s'ils pensaient qu'aucune conclusion valide ne s'ensuivait), une feuille leur était présentée sur laquelle toutes les autres conclusions possibles étaient nommées, et ils devaient alors indiquer celles qu'ils avaient considérées – en plus de leur réponse. Parmi les neuf conclusions présentées pour chaque syllogisme, en moyenne les participants n'en considérèrent qu'une supplémentaire (1,09). Il s'agit là d'un nombre très faible d'alternatives considérées, et deux éléments indiquent qu'il ne s'agit même pas de réelles tentatives de falsifier la conclusion.

Le premier est qu'il n'y a aucune corrélation entre le nombre d'alternatives considérées et le taux de réponses correctes. Si les rares alternatives considérées l'étaient bien dans le but de falsifier la conclusion, elles devraient permettre de

rejeter davantage de conclusions non nécessaires (tout en ne rejetant pas les conclusions nécessaires que la construction de modèles alternatifs ne peut que laisser intactes). De plus, le nombre d'alternatives considérées était totalement indépendant de la difficulté et du degré d'indétermination des syllogismes présentés. Or, selon la théorie des modèles mentaux, c'est dans les cas de forte indétermination que la procédure de falsification de la conclusion devrait forcer les participants à envisager d'autres hypothèses. Il paraît plus probable donc qu'il faille chercher ailleurs pourquoi dans certains cas les participants considèrent plusieurs conclusions. Une possibilité est la faiblesse et l'ambiguïté des intuitions initiales. Etant donné le caractère très abstrait de la tâche (rappelons que les prémisses sont faites sur le modèle suivant : « All of the buskers are computer operators. None of the computer operators are boxers », ce qui ne risque guère de motiver les participants), il est fort probable qu'ils n'aient que des intuitions très faibles sur la réponse, ou même que plusieurs intuitions faibles soient en compétition rapprochée. Dans ce cas, il est normal que les participants examinent ces différentes options, sans toutefois nécessairement chercher à les falsifier : ils peuvent simplement chercher laquelle serait la plus facile à défendre.

Deux autres expériences menées à la suite de celle-ci peuvent être interprétés comme confirmant cette explication (Newstead et al., 1999, expériences 2 et 3). Dans ces expériences, avant de donner la conclusion des syllogismes, les participants devaient dessiner des formes (les syllogismes impliquaient cette fois des formes ayant différentes caractéristiques tels que des carrés avec des bords épais et des rayures) correspondant à la situation décrite par les prémisses. Ils étaient fortement encouragés à construire jusqu'à trois représentations différentes. Si ces représentations étaient ensuite utilisées pour falsifier la conclusion envisagée, leur diversité devrait corréler avec la réussite à la tâche. A l'inverse, si ces représentations ne font que refléter des intuitions différentes que les participants peuvent se former sur la base des prémisses, alors il ne devrait pas y avoir un tel effet. On pourrait même s'attendre, en extrapolant un peu, à ce que le fait de construire plusieurs représentations signifie que plusieurs conclusions seront examinées, ce qui renforce les chances (déjà élevées) que leur examen ne soit que superficiel et n'implique pas de tentative de falsification. Tous les résultats indiquèrent une corrélation négative (mais non significative, bien que dépassant les – 0,40 dans un cas) entre la diversité des représentations créées et la réussite à la tâche, ce qui montre bien que lorsque

plusieurs représentations sont créées, elles ne sont pas utilisées pour tenter de falsifier la conclusion.

Une autre expérience, de plus grande ampleur quant à l'éventail des syllogismes utilisés, a conduit ses auteurs à une conclusion également pessimiste quant à la tendance naturelle des participants à chercher des contre-exemples. Dans cette expérience, les participants étaient confrontés à des syllogismes dont la conclusion était donnée, et ils devaient déterminer si elle était nécessaire (dans une condition), ou possible (dans l'autre condition) (Evans et al., 1999). Dans le cadre présenté ici, on peut s'attendre à trouver au moins une différence entre ces deux conditions. Pour ce qui est de la condition nécessaire, la théorie argumentative prédit qu'ils se contenteront de chercher des arguments soutenant leur intuition initiale. Or leur intuition initiale est justement due à leur représentation des prémisses : ces prémisses peuvent dès lors constituer, à leurs yeux, un argument suffisant pour soutenir leur intuition. Dans un cadre similaire, c'est exactement le comportement que j'ai observé chez de nombreux participants à qui je demandais de justifier leurs réponses dans le cas des problèmes du CRT comme le 'bat and ball' (voir section 5.2.1) : ils se contentaient souvent de répéter les termes du problème. Etant donné que les termes du problème évoquent chez eux une forte intuition pour une réponse donnée, ils les considèrent comme des arguments suffisants pour soutenir la conclusion qu'ils défendent. On peut très bien imaginer que les participants confrontés à des tâches de syllogismes se comportent de la même manière : une fois qu'ils ont l'intuition que la conclusion découle (ou ne découle pas) des prémisses, ils peuvent se contenter de ces mêmes prémisses comme argument, ce qui rend caduque toute recherche supplémentaire. Cette interprétation est cohérente avec les résultats de cette étude : « This suggests that any search for counterexample models is weak in the present study and that most participants are basing their conclusions on the first model that occurs to them » (p.1505)[47].

Lorsque les participants doivent se prononcer sur le fait que la conclusion est possible ou non, par contre, ils doivent avoir recours à une stratégie quelque peu différente. Lorsqu'ils ont l'intuition que la conclusion est possible, il ne devrait pas y

[47] Au crédit des participants de cette expérience : ils avaient à résoudre 64 syllogismes, et ce juste après avoir résolu 28 inférences immédiates. Il n'est guère surprenant qu'après plus de 50 problèmes de logique abscons on n'ait plus guère envie de raisonner et qu'on se contente de peu pour ce qui est de défendre notre réponse, si on tente même encore de le faire.

avoir de différence avec la condition dans laquelle ils doivent se prononcer sur sa nécessité : si elle est nécessaire, elle est forcément possible – le test qui devrait être rigoureux dans la condition nécessaire ne l'est déjà pas assez, il n'y a aucune raison qu'il le soit plus dans la condition possible. Par contre, les choses changent lorsque les participants ont l'intuition qu'il faut rejeter la conclusion. Dans ce cas, ils ne doivent pas simplement trouver des arguments qu'ils jugent suffisants pour montrer que la conclusion ne découle pas des prémisses (ce à quoi les prémisses mêmes pourraient suffire), mais ils doivent trouver des arguments montrant qu'elle est impossible – une condition plus forte que non nécessaire. Dans cette situation, la réaction naturelle est de chercher des possibilités alternatives de parvenir à la conclusion. Si cette recherche n'est pas fructueuse, il s'agit d'un argument pour montrer que la conclusion est bien impossible. Il faut souligner que cette recherche n'a pas pour but de montrer que l'intuition initiale du participant est mauvaise, mais au contraire qu'elle est bonne. C'est le seul cas dans lequel le participant doit activement chercher des solutions alternatives – à nouveau, non pas pour falsifier sa réponse initiale que la conclusion est impossible, mais bien pour la soutenir. C'est précisément ce que montrent les résultats.

Les syllogismes pertinents pour montrer ce phénomène sont ceux dont la conclusion est possible mais que les participants de la condition nécessaire rejettent : on peut interpréter ce rejet dans la condition nécessaire comme un signe que l'intuition initiale des participants tend au rejet de la conclusion, ce qui devrait toujours être le cas dans la condition possible. Dans cette condition, les participants devraient alors chercher des possibilités alternatives et, étant donné qu'elles existent (il s'agit de syllogismes dont la conclusion est possible), certains d'entre eux au moins devraient les trouver, et donc accepter, finalement, la conclusion. Le taux d'acceptation devrait donc être plus élevé dans la condition possible que dans la condition nécessaire. C'est en effet ce qui est observé : dans l'expérience 3, qui visait spécifiquement ces problèmes, les conclusions possibles qui tendent à être rejetées dans la condition nécessaire étaient acceptées deux fois plus souvent dans la condition possible (38%) que dans la condition nécessaire (19%). On peut également savoir que cet effet n'était pas dû à un laxisme généralisé dans la condition possible : le taux d'acceptation n'y était supérieur que dans le cas qui vient d'être décrit – pour tous les autres types de syllogismes, il n'y avait pas de différence significative entre les deux conditions. En particulier, dans le cas des syllogismes dont la conclusion est

réellement impossible, la recherche de possibilités alternatives de parvenir à la conclusion ne pouvait que s'avérer vaine, et les participants ne furent donc logiquement pas significativement plus nombreux à accepter les conclusions impossibles dans la condition possible que dans la condition nécessaire. Les auteurs en concluent que : « Thus we have clear evidence that people search for alternative models to prove the possibility of conclusions that are not supported by the first model considered, but in this case less clear evidence that people search for counterexamples to establish the necessity of conclusions that are supported by the first model considered. » (p.1507).

On peut conclure de cette étude, ainsi que des résultats des expériences de Newstead et al. (1999) que les participants, dans leur très grande majorité, ne cherchent pas à falsifier leur réponse initiale. Ils ne cherchent que très rarement des contre-exemples, et lorsqu'ils considèrent spontanément des conclusions alternatives, cela ne fait que refléter le fait qu'ils ont plusieurs intuitions sur la réponse possible. Enfin, dans le seul cas où ils cherchent activement des réponses alternatives, ce n'est pas dans une tentative de falsification, mais bien pour montrer que leur réponse est correcte.

Syllogismes : biais de croyance

Un des phénomènes les plus étudiés dans le cadre des syllogismes est le biais de croyance. Il s'agit de la tendance à prendre en compte la crédibilité de la conclusion dans la détermination de la validité logique de l'argument, alors même qu'elle ne devrait pas l'être (voir section 3.1 pour des exemples). Comme expliqué précédemment, il y a deux phénomènes principaux et assez robustes : d'une part un effet de la crédibilité (le biais en lui-même) qui fait que les conclusions crédibles (valides et non valides) tendent à être davantage acceptées que les conclusions non crédibles (valides et non valides). Le second effet est une interaction : alors que les syllogismes aux conclusions crédibles tendent à être acceptés qu'ils soient logiquement valides ou non, il y a un effet de la validité beaucoup plus fort lorsque la conclusion est non crédible. Malheureusement, les résultats sur le biais de croyance sont extrêmement difficiles à interpréter car de nombreux facteurs, au-delà de la crédibilité de la conclusion et de la validité du syllogisme, entrent en jeu. Par

exemple, les réactions sont différentes selon la difficulté du syllogisme (en termes de modèles mentaux, le nombre de modèles dont la construction est nécessaire pour sa résolution) et selon les prémisses utilisées. Il n'est pas toujours facile de créer des raisonnements valides soutenant des conclusions non crédibles. Cela peut impliquer des prémisses douteuses, des prémisses clairement fausses, ou des prémisses contenant des non-mots : selon le type de prémisses, les résultats peuvent être différents (voir par exemple V. A. Thompson, 1996, sur la question du statut des prémisses, et Klauer et al., 2000, plus généralement). Je ne tenterai donc pas ici d'établir un modèle détaillé du comportement des participants dans ce type de tâche, et me contenterai de noter les conclusions d'une étude de grande ampleur faite sur le sujet.

Dans cette étude, Klauer et ses collègues se sont non seulement livrés à une méta-analyse des données de la littérature, mais ont également conduit huit nouvelles expériences afin de déterminer les facteurs influant la résolution de syllogismes en fonction de la crédibilité de la conclusion (Klauer et al., 2000). Leurs conclusions les plus générales rejoignent celles des autres études portant sur les syllogismes : les participants ne construisent qu'un seul modèle des prémisses et de la conclusion, et ne cherchent que rarement des modèles alternatifs (hypothèse 2, p.871). Lorsque la conclusion n'est pas crédible, ils procèdent cette fois à un *test négatif* (dans les termes des théories du test d'hypothèse évoquées plus haut). Dans le cas des syllogismes, un test négatif correspond à tester la négation logique de la conclusion donnée. Par exemple :

Tous les poissons sont des acquérites
Tous les acquérites sont des truites
Donc tous les poissons sont des truites

Dans ce cas, un test négatif consiste à essayer de former un modèle intégrant les prémisses et la négation logique de la conclusion, à savoir « certains poissons ne sont pas des truites ». Les auteurs laissent cependant un aspect non éclairci, aspect qui s'était avéré crucial pour ce qui est du test d'hypothèse : une fois que les participants s'engagent dans ce test négatif, le font-ils dans une réelle optique de falsification, ou dans une optique de vérification ? Pour le savoir, il faudrait connaître les attentes des participants : pensent-ils qu'ils vont réussir à créer un modèle leur permettant de

soutenir « certains poissons ne sont pas des truites » ? Si c'est le cas, cela signifie qu'ils sont bien en train d'essayer de falsifier la conclusion originale (et non crédible). Il serait extrêmement surprenant que ce ne soit pas le cas : « certains poissons ne sont pas des truites » fait partie de leur base de connaissance, et on voit mal comment ils pourraient s'attendre à être dans l'impossibilité de trouver des arguments pour quelque chose de trivialement vrai. Il s'agirait donc là d'une réelle tentative de falsification, les participants ne finissant par accepter la conclusion non crédible comme étant valide que s'ils échouent à trouver un modèle leur permettant de soutenir sa négation logique. Cependant, dans cet unique cas dans lequel la falsification est commune, l'énoncé n'a pas été généré par les participants : au contraire, leur première intuition est une intuition de rejet.

Avant de conclure sur les syllogismes en général, on peut mentionner une étude plus récente qui prétend remettre en cause ces conclusions sur le biais de croyance. Au moyen d'une nouvelle expérience, Thompson et ses collègues attaquent la thèse selon laquelle les participants ne raisonneraient pas du tout lorsque la conclusion d'un syllogisme est crédible, mais raisonneraient au contraire afin de falsifier les conclusions non crédibles (V.A. Thompson, Striemer, Reikoff, Gunter, & Campbell, 2005). La première partie de leur expérience consistait en la résolution normale de syllogismes, les participants étant chronométrés. Les temps de réaction furent légèrement plus longs pour les syllogismes aux conclusions crédibles (21s versus 20s, une différence qui, bien que significative, ne me paraît pas révélatrice). Cela montre en tout cas que les participants réfléchissent aussi lorsque la conclusion est crédible. Ce résultat s'accorde cependant très bien avec la théorie argumentative selon laquelle les participants vont tenter de justifier leur réponse dans tous les cas.

Dans la seconde phase de l'expérience (après avoir donné leurs réponses), les participants devaient dessiner des diagrammes correspondant aux prémisses. Pour la théorie des modèles mentaux, par exemple, le raisonnement passe par la construction de modèles alternatifs. Donc, si les participants raisonnent plus lorsque les conclusions sont non crédibles, ils devraient dessiner plus de diagrammes dans ce cas. Or aucune différence ne fut observée dans le nombre de diagrammes dessinés, qu'il s'agisse de syllogismes à la conclusion crédible ou non crédible. Mais dans le cadre de la théorie de Klauer et collègues, cela n'est pas surprenant. Confrontés à une conclusion non crédible, les participants ne cherchent à construire qu'un seul modèle intégrant prémisses et conclusion dans tous les cas. La seule différence est que dans

le cas des conclusions non crédible, le modèle tente en fait d'intégrer la négation logique de la conclusion avec les prémisses, et non la conclusion elle-même. Il n'y a dès lors pas de raison de s'attendre à ce que la représentation des prémisses soit plus complexe dans un cas que dans l'autre. On peut donc dire que ce résultat vient au contraire confirmer le fait que les participants ne se représentent dans tous les cas qu'un seul modèle des prémisses : en moyenne, ils ne furent capables de dessiner qu'un seul diagramme correct, et ce quelque soit le type de syllogisme.

Finalement, un autre résultat vient confirmer l'interprétation selon laquelle les participants ont une intuition initiale d'acceptation des conclusions crédibles et de rejet des conclusions non-crédibles, intuitions qu'ils ne font que chercher à justifier par la suite : lorsque les participants doivent répondre rapidement, ils ont une forte tendance à se contenter d'accepter les syllogismes aux conclusions crédibles et de rejeter ceux aux conclusions non crédibles (Evans & Curtis-Holmes, 2005).

On peut conclure cette partie sur les syllogismes sur les trois points suivants, qui concordent bien avec les prédictions de la théorie argumentative. Tout d'abord, et il s'agit du résultat le plus important, les participants ne tentent que très rarement de falsifier leur intuition initiale. Deuxièmement, le seul cas dans lequel ils s'engagent régulièrement dans une vraie stratégie de falsification est lorsque leur intuition initiale implique le rejet de la conclusion car elle n'est pas crédible. S'il s'agit bien d'une stratégie de falsification, elle ne vise pas à falsifier l'intuition initiale, mais bien à la confirmer – c'est la conclusion non crédible que les participants cherchent à falsifier. Enfin, les participants sont capables de reconnaître que leur intuition initiale ne peut pas être adéquatement soutenue et doit donc être abandonnée. Cela arrive au moins dans les deux cas suivants. Lorsqu'ils pensent initialement qu'une conclusion est impossible et que, cherchant à prouver qu'elle est impossible en montrant qu'on ne peut pas la dériver des prémisses, ils trouvent en fait un moyen de la dériver et en viennent donc à l'accepter. Et lorsque, cherchant un modèle soutenant la négation logique d'une conclusion non crédible, ils n'en trouvent pas et en viennent donc à accepter la validité de la conclusion. Le raisonnement des participants est donc biaisé car ils ne font presque que chercher à confirmer leurs intuitions initiales, mais s'ils ne peuvent la justifier, ils préfèrent donner une autre réponse – une conclusion que l'on retrouvera dans le cas du raisonnement motivé.

6.4 Conclusion : un biais métareprésentationnel

Pour conclure cette partie sur le biais de confirmation en psychologie du raisonnement, j'aimerais revenir sur l'importance de la différence entre mécanismes représentationnels 'simples' et mécanismes métareprésentationnels. En particulier, il me semble important de souligner que certaines stratégies peuvent être efficaces dans un cas, mais dangereuses (épistémiquement tout au moins) dans l'autre. Prenons l'exemple du test d'hypothèse. Vous formez une hypothèse sur la bienveillance de Jean : disons que vous le pensez plutôt gentil. Vous pouvez observer son comportement afin de voir s'il se conforme à votre hypothèse. Peut-être certains biais surviendront-ils dans l'appréciation que vous en ferez, mais il n'empêche que le comportement de Jean est globalement une variable externe, qui n'est pas liée à l'hypothèse que vous testez. Si vous ne faites que l'observer, sans interagir avec lui, il n'y a pas de raison que votre hypothèse influence son comportement (voir la section suivante pour des effets de recherche sélective d'information). Ses actes constituent dès lors un bon moyen de tester votre hypothèse : il y a une réelle chance qu'elle soit falsifiée par exemple.

Imaginez maintenant que vous testiez cette même hypothèse au moyen d'un mécanisme métareprésentationnel. Par exemple, vous essayez de chercher en mémoire des éléments pertinents par rapport à votre hypothèse. Dans ce cas, les mécanismes de pertinence peuvent créer des biais importants : si vous pensez 'est-ce que Jean est gentil ?', les souvenirs qui sont rendus les plus pertinents sont ceux dans lesquels Jean a été, en effet, gentil. Cet effet est dû à la façon dont s'organisent les représentations : afin qu'on puisse s'y retrouver dans nos souvenirs, il est nécessaire de les ordonner. Dans ce cas, la recherche risque de sur-représenter très fortement les cas dans lesquels Jean s'est bien comporté[48]. A nouveau, un tel phénomène ne risque guère de survenir lorsque vous observez le monde directement : alors que vos représentations des actions bienveillantes de Jean sont groupées en mémoire, ses

[48] Il ne s'agit là que de la partie bottom-up (ou efforts) de la pertinence. On peut également imaginer que des réponses positives ou négatives à la question « Jean est-il gentil ? » soient plus ou moins pertinentes (partie top-down, ou effets), et cela pourrait également influencer la recherche. Mais à nouveau, cette influence peut mener à des distorsions indues.

actions bienveillantes (dans le monde) ne le sont pas nécessairement, et le simple fait que vous y pensiez n'augmente pas les chances qu'elles se produisent.

C'est pour cette raison que l'absence de falsification spontanée observée dans les expériences qui viennent d'être relatées peut s'avérer dommageable (voir Matthew & Schrag, 1999, pour une démonstration formelle). Ne pas chercher spécialement à falsifier ses hypothèses lorsque c'est le monde qui les teste n'est pas nécessairement un problème majeur : si Jean est un vrai salaud, nous nous en rendrons probablement compte assez vite – les chances que nous n'assistions qu'à de bonnes actions de sa part sont minces. Par contre, lorsqu'il s'agit de chercher parmi nos propres représentations, il peut s'agir d'un biais sérieux : il y a un risque réel de former des opinions fortement influencées par les hypothèses que nous avons considérées en premier lieu. Or c'est précisément cette absence de falsification que l'on peut observer dans toutes les tâches qui viennent d'être passées en revues – tests d'hypothèses, tâche de sélection ou syllogismes. Imaginez un animal se demandant si un endroit est sûr, s'il ne contient pas de dangers. Admettons que son intuition première soit de penser qu'il l'est. S'il se comportait comme les participants de ces expériences, il chercherait alors à justifier cette intuition, par exemple en se rappelant toutes les fois ou il y est allé sans qu'il ne rencontre de prédateurs. Cela ne pourrait avoir que pour effet de le renforcer dans son intuition que cet endroit est, en effet, sûr, quitte à l'amener à s'aventurer plein de confiance sur un territoire potentiellement dangereux (voir la section sur la sur-confiance, la polarisation et le renforcement). Dans ce cas, on peut estimer que l'intuition initiale est formée de façon assez optimale (sur la base des visites précédentes de cet endroit), et que toute déviation ne peut entraîner que des baisses de performances.

Pour la théorie argumentative, il s'agit là du fonctionnement normal du raisonnement : son objectif n'est pas de nous aider à former des croyances plus exactes, mais bien de trouver des raisons pouvant soutenir certaines conclusions. En ne cherchant pas à falsifier les hypothèses, ou intuitions, des participants, et au contraire en cherchant à les justifier, il ne fait donc que son travail. On peut cependant se poser deux questions : si ses effets sont nuls, ou potentiellement néfastes, pourquoi le raisonnement est-il utilisé dans ces circonstances ? Et, si les mécanismes de tests d'hypothèse utilisés sont efficaces lorsqu'il s'agit du rapport entre de simples représentations et le monde, ne peut-on expliquer ces résultats comme une simple utilisation mal à propos de mécanismes bien adaptés par ailleurs ?

Pour répondre à la première question, il faut se rappeler que les participants sont placés dans une situation qui promeut l'utilisation du raisonnement pour les deux raisons suivantes (voir les prédictions faites au tout début de cette partie). D'une part ils n'ont pas envie de passer pour des incompétents : ils savent que leurs réponses vont être examinées, cela ressemble à un test scolaire, ou d'intelligence, il y a donc une certaine pression sociale pour bien réussir, pression que les conditions d'anonymat supposé ne parviennent jamais à lever totalement. D'autre part, leurs intuitions sur la réponse correcte seront souvent faibles. Dans l'incertitude vis-à-vis de ce qui constitue la bonne réponse, ils pourront donc chercher à les départager en choisissant celle qui se justifie le plus aisément. Deux conditions prévues par la théorie argumentative pour l'utilisation du raisonnement en situation de prise de décision sont donc remplies. Ceci ne fait cependant que répondre à l'aspect 'proximal' de la question : il s'agit là, selon la théorie, des mécanismes qui font que le raisonnement est déclenché dans ce type de tâche. La réponse au versant 'ultime' de cette question a déjà été abordée, elle aussi, au début de cette partie : si le raisonnement est activé dans ces circonstances, c'est pour que, alors que nous risquons de prendre une décision qui pourrait être jugée comme étant mauvaise, nous soyons au moins capables de la justifier. Et dans la mesure où les participants sont généralement capables de donner ne serait-ce qu'un semblant d'explication, de justification pour leur choix (même s'il ne correspond parfois qu'à une répétition des prémisses), le raisonnement remplit sa fonction.

Pour ce qui est de la seconde question, on peut lui adresser une réponse qui est récurrente lorsqu'on se place dans un cadre évolutionniste (ou dans un cadre d'analyse rationnelle tel que celui d'Anderson [1991] ou de Marr [1982]). S'il faut choisir entre une théorie qui explique les données par un défaut, par une contrainte, et une théorie qui prédit ces mêmes données comme résultant du bon fonctionnement d'un mécanisme dont on peut expliquer pourquoi il se comporte de cette façon, il convient (toutes choses égales par ailleurs) de favoriser la seconde. Or, dans ce cas, une utilisation inappropriée de mécanismes adaptés à un certain niveau constitue bien un tel défaut : pourquoi ces mécanismes sont-ils utilisés s'ils ne sont pas adaptés ? Pourquoi ne sont-ils pas corrigés ? Les réponses risquent alors de s'exprimer en termes de contraintes : car une telle adaptation serait trop coûteuse… mais alors pourquoi utiliser ces mécanismes mal adaptés en premier lieu ? La faiblesse générale de ce type d'explication ayant déjà été abordée dans la section sur la comparaison

entre les théories à processus duel classiques et la théorie argumentative, je n'y reviendrai pas davantage ici.

6.5 Recherche sélective d'informations

Le plus souvent, on ne veut savoir que pour en parler.
Blaise Pascal, Pensées, XXIV

Nous venons de voir que les biais dans la recherche d'arguments pouvaient s'avérer dommageable d'un point de vue épistémique. Il n'est donc guère surprenant que ces biais soient tenus pour responsables d'actions aux conséquences néfastes (voire dramatiques) : comment bien agir si on se fait des idées systématiquement biaisées sur le monde ? Alors que la majorité des travaux qui viennent d'être passés en revue étaient tirés de la psychologie du raisonnement, une immense tradition de recherche, portant sur la recherche sélective d'informations, pointe dans la même direction. La recherche sélective d'informations ('selective exposure') est la tendance à « rechercher des informations qui étayent nos opinions et éviter les informations qui les contredisent » (S. M. Smith, Fabrigar, & Norris, 2008, p.464). Dans leur revue de la littérature sur le sujet, Smith et al. (2008) décrivent les débuts prometteurs de cette idée, rapidement suivis d'un revirement entraîné par quelques résultats négatifs, avant que, dans le courant des années 1960, de nouvelles expériences démontrent la réalité du phénomène. Depuis, une montagne de travaux s'est accumulée, des dizaines de facteurs modérateurs possibles ayant été étudiés. Il n'est pas possible de passer ces travaux en revue ici, même succinctement. Une expérience récente servira d'illustration de recherche sélective d'information puis, étant donné qu'il semble maintenant difficile de contester la réalité même du phénomène, des considérations théoriques sur ses sources et ses effets seront discutées.

Dans une manipulation très simple, Brannon et ses collègues, après avoir évalué les attitudes des participants, leurs dirent qu'ils pourraient plus tard participer à une expérience dans laquelle ils devraient lire des articles de journaux (Brannon, Tagler, & Eagly, 2007). Une liste de titres d'articles (accompagnés de leurs résumés

dans une seconde expérience) leur fut ensuite présentée, et ils devaient noter s'ils trouvaient désirable de lire chaque article. Les résultats montrèrent que les articles dont les titres (et / ou les résumés) étaient cohérents avec les attitudes des participants étaient préférés en tant que lectures potentielles, et que ce résultat était modéré par la force et l'extrémité des attitudes des participants. Il s'agit là d'un des exemples les plus simples de résultat démontrant une recherche sélective d'information, et il est donc important de rappeler que de nombreuses autres méthodes, expérimentales ou de terrain, ont également permis de démontrer l'existence de ce phénomène.

Pour intuitifs qu'ils paraissent, ces résultats soulèvent néanmoins des questions théoriques intéressantes. En particulier, il est clair qu'un tel mode de recherche d'information aura souvent des conséquences néfastes. En effet de telles conséquences ont été observées à de nombreuses reprises : décisions médicales (Nemeth & Rogers, 1996), maintenance indue de stéréotypes (Cameron & Trope, 2004; Johnston, 1996), prise de risque potentielle avec des partenaires sexuels (Hennessy, Fishbein, Curtis, & Barrett, 2008), ou encore résultats sous optimaux dans des négociations (Pinkley, Griffith, & Northcraft, 1995) – et il faut garder à l'esprit qu'il ne s'agit là que d'une liste très partielle se concentrant sur les démonstrations récentes. On ne sera donc guère surpris d'apprendre que les animaux ne semblent pas être sujets à ce type de biais, et se comportent de manière beaucoup plus optimale que nous (voir par exemple Real, 1992, chez les bourdons, et Dall, Giraldeau, Olsson, McNamara, & Stephens, 2005, plus généralement). Avant d'en venir à l'explication dans les termes de la théorie argumentative, on peut se pencher sur d'autres explications qui ont été avancées pour expliquer l'aspect 'adaptatif' de la recherche sélective d'information.

Ainsi, Jonas et ses collègues avancent les deux bénéfices suivants pour le biais de confirmation (qui était observé sous forme de recherche sélective d'information dans ce cas) : « support the decisionmaker's ability to act [and] facilitates an emphatic implementation of the decision »[49] (Jonas, Greenberg, & Frey,

[49] Ce qui me rappelle cette affiche qui était placée au dessus de la machine à café de l'institut : « Coffee: do stupid things faster and with more energy ».

2003, p.1187). On peut considérer qu'il ne s'agit en fait que de deux versants d'un même phénomène : si on a plus de confiance dans une décision (que ce soit parce qu'on a recueilli des informations biaisées à son sujet ou pour d'autres raisons), on a plus de chance de la prendre, et on risque également d'y investir plus d'énergie. Un premier réflexe est de se dire que cet argument est mauvais : dans tous les cas, il vaut mieux agir sur la base d'informations qui ont le plus de chances possible d'être vraies. Si on peut considérer cela comme un truisme, on ne peut dériver de cette prémisse la conclusion que les mécanismes qui cherchent et traitent les informations ne doivent pas être biaisés, qu'ils doivent toujours être le plus objectif possible[50]. Cependant, même s'il est toujours préférable d'avoir raison, toutes les erreurs n'ont pas le même coût. Il est donc possible qu'un mécanisme moins fiable (dont les résultats sont moins souvent en concordance avec la réalité), mais dont les erreurs ne sont que peu coûteuses soit sélectionné au détriment d'un mécanisme plus fiable mais dont les erreurs sont plus coûteuses.

Un exemple est celui des nourritures toxiques. Imaginons un mécanisme dont la fonction est de déterminer si un champignon est vénéneux ou non. La version A du mécanisme a un taux de faux négatifs de 5%, et un taux de faux positifs de 1% : dans 5% des cas, il prend un champignon vénéneux pour un champignon comestible (coûts potentiellement très importants) et dans 1% des cas il commet l'erreur inverse. La version B a quant-à-elle un taux de faux négatifs de 1% mais un taux de faux positifs de 10%. Elle est donc dans l'ensemble moins fiable que la version A. Cependant, les erreurs qu'elle commet sont beaucoup moins coûteuses : mieux vaut délaisser un champignon comestible que manger un champignon vénéneux. Si la proportion de champignons comestibles et vénéneux est approximativement égale, c'est la version B, la moins fiable pourtant, qui sera favorisée.

On pourrait imaginer qu'un tel raisonnement explique certaines formes d'optimisme. De même qu'on pourrait considérer la version B comme une version 'pessimiste' (elle voit du poison là ou il n'y en a pas), si la balance des coûts est inversée, c'est une version 'optimiste' qui prévaudra. Dans quelle situations est-ce qu'il peut-être moins coûteux d'agir que de s'abstenir d'agir ? Un domaine dans lequel ces circonstances peuvent régulièrement émerger est le domaine social. Pour

[50] Cette discussion pourrait presque être une transcription d'un échange avec Dan Sperber, qui m'avait convaincu justement de ne pas tirer cette conclusion de cette prémisse.

embarrassantes qu'elles soient, la plupart des 'gaffes' que nous pouvons commettre en société ont des conséquences relativement légères : nous ne nous faisons pas dévorer par un tigre, nous ne chutons pas dans un précipice, nous n'ingérons pas de nourriture toxique. On peut donc raisonnablement s'attendre à ce que, dans le domaine social (ou, au moins, certaines parties de ce domaine), les gens tendent à être biaisés pour interpréter les informations d'une façon qui les pousse à agir.

Ces considérations sont tout aussi valables pour ce qui est des biais dans la façon dont les informations sont traitées que dans la façon dont elles sont recherchées. De même qu'un signe ambigu qu'un champignon est vénéneux devrait plutôt être interprété comme un signe qu'il est, en effet, toxique, si le choix est donné entre une information qui tendra à indiquer qu'il est vénéneux et une information qui tendra à indiquer l'inverse, mieux vaut choisir la première. Dans le domaine social, cela pourrait se traduire par une affinité avec les personnes qui tendent à nous renvoyer une image positive de nous-mêmes.

Il est possible que ce type d'explication permette de rendre compte d'un certain nombre de résultats montrant une recherche sélective d'information. Ainsi, les participants d'une expérience étaient plus motivés pour lire des informations discréditant les tests d'intelligence s'ils pensaient avoir juste obtenu de mauvais résultats à un tel test (Frey & Stahlberg, 1986). Dans ce cas, il pourrait s'agir d'une manifestation d'un biais général nous faisant rechercher des informations positives sur nous-mêmes (voir également Holton & Pyszczynski, 1989).

La grande majorité des travaux sur la recherche sélective d'information s'inscrivent cependant dans la tradition d'études sur la dissonance cognitive. Cela signifie que le déterminant primaire de la recherche d'information n'est pas le fait qu'elles nous donnent une image positive de nous, ou qu'elles inspirent l'optimisme plus généralement, mais le fait qu'elles soient cohérentes, compatibles avec nos croyances ou attitudes préalables. Cette caractéristique des informations est orthogonale à celles qui viennent d'être envisagés. Les biais dont je parlais dans le paragraphe précédent sont spécifiques à un domaine particulier, ils dépendent des coûts relatifs des erreurs, et d'autres paramètres propres à ce domaine. Le fait de rechercher des informations cohérentes avec nos attitudes préalables peut s'appliquer à tous les domaines. Dans certains cas, ces deux phénomènes pourront avoir des

effets similaires – ça sera par exemple le cas si nous avons une attitude positive vis-à-vis de nous-mêmes – et dans d'autres cas les effets pourront être opposés.

Si nous avons pu trouver une justification normative pour l'existence du premier type de biais, cela me paraît beaucoup plus difficile dans ce deuxième cas. De manière générale, une information est valable, intéressante, dans la mesure où elle est surprenante (Shannon, 1948). Une information n'a d'effet que si elle nous force à réviser nos croyances. A la limite, si elle n'a pas du tout d'effet, il ne s'agit même plus d'une information. Un mécanisme dont la fonction est d'acquérir des informations devrait donc être renforcé dans la mesure où les informations qu'il acquiert sont valables, où elles changent nos croyances[51]. Ceci est vrai indépendamment de la valence du contenu des informations. Par exemple, un mécanisme dont la fonction serait de recueillir des informations sur la fidélité de notre partenaire doit être récompensé s'il nous apprend de nouvelles informations, même si ces informations elles-mêmes peuvent nous rendre malheureux. Il est donc réellement surprenant que l'on observe le goût inverse – un goût pour l'absence d'information pourrait-on dire – dans tant de situations. Dans le cadre de la théorie de la dissonance cognitive, ceci est généralement expliqué par le fait que les informations consonantes sont plaisantes, ou qu'elles requièrent moins d'effort de traitement (voir Smith et al., 2008, p.479, et les références données). Quelque soit la validité de ces arguments, ils ne répondent à la question que du point de vue proximal. En particulier, il faudrait savoir pourquoi les informations cohérentes ont une valeur hédonique alors que ce sont plutôt les informations incohérentes qui devraient en avoir (dans la mesure où elles sont généralement plus informatives).

[51] En fait, cet argument devrait être un peu plus compliqué que ceci. Entre deux mécanismes qui informent sur une même source, et toutes choses égales par ailleurs, un mécanisme qui donne plus d'informations ne devrait pas être favorisé. Car si deux mécanismes sont aussi fiables, le fait que l'un donne plus d'informations que l'autre signifie simplement qu'il fait plus d'erreurs dans le sens de penser voir un changement là ou il n'y en a pas que d'erreurs inverses (et vice versa pour le mécanisme qui donne moins d'informations). Par contre, si les deux mécanismes donnent des informations sur des domaines différents, également importants, mais dont la variance est inconnue, alors les mécanismes qui donnent plus d'informations devraient être récompensés car cela signifie que la variance du domaine qu'ils renseignent est plus importante et donc qu'ils sont plus utiles.

La théorie argumentative explique ces effets par la pertinence qu'ont les informations consonantes en tant qu'arguments potentiels[52]. Si on prend l'exemple de l'expérience qui a été relatée plus haut, les participants devaient choisir des arguments pour ou contre l'avortement (l'IVG). Si l'objectif des participants était de se former la meilleure opinion possible sur le sujet, ils devraient plutôt choisir les arguments s'opposant à leur point de vue – on peut raisonnablement penser qu'il y a plus de chances qu'il s'agisse d'arguments nouveaux, et donc plus informatifs. Par contre, si le contenu des articles est pertinent en tant qu'arguments potentiels soutenant notre point de vue, alors il est normal de préférer ceux qui sont justement cohérents avec notre point de vue. Il s'agit d'un argument similaire à celui qui avait été fait pour expliquer les résultats dans certaines versions de la tâche de sélection : les participants y étaient motivés pour retourner certaines cartes car elles pourraient fournir des arguments contre une règle qu'ils désiraient rejeter. Le phénomène de recherche sélective d'information rejoint donc les résultats passés en revue plus haut sur le biais de confirmation et montrent que le raisonnement peut amener à donner des réponses non normatives car, plutôt que d'évaluer les informations objectivement, il cherche des arguments.

[52] Comme le remarquent Smith et al. (2008), il est également possible que les participants anticipent leur réaction de contre-argumentation naturelle face à des informations dissonantes, mais ici encore comment expliquer ce besoin de contre-argumenter autrement que dans un cadre argumentatif (imaginez un lapin en train d'essayer de se persuader que la forme rouge orangé avec la grande queue touffue n'est pas un renard car cela n'est pas cohérent avec ses croyances…).

7 Le raisonnement motivé

So convenient a thing it is to be a reasonable creature, since it enables one to find or make a reason for everything one has a mind to do.

Benjamin Franklin, Autobiographie

Nous venons de voir que même dans les tâches de psychologie du raisonnement, dans lesquelles on pourrait penser justement que le raisonnement serait employé assez naturellement, il l'est plus pour chercher à justifier les décisions des participants quant aux réponses à donner que pour réellement évaluer les arguments présentés. Dans ce chapitre, nous allons nous intéresser à d'autres circonstances dans lesquelles le raisonnement est utilisé d'une façon similaire. Il s'agit de situations dans lesquelles les participants sont motivés pour parvenir à une conclusion et cherchent des arguments pour la soutenir. En raison de cet aspect motivationnel, ces travaux sont souvent étiquetés comme portant sur le '*raisonnement motivé*'. Le début de ce chapitre sera consacré à une revue des travaux qui se situent directement dans cette tradition, puis nous passerons à d'autres recherches qui, s'en inspirant plus ou moins directement, présentent des conclusions similaires. Enfin, nous conclurons en notant un point de discorde potentiel entre la théorie argumentative et certaines théories du raisonnement motivé, puis divers éléments correspondant aux prédictions de la première seront présentés.

7.1 La théorie de Kunda

A la suite de la revue souvent citée de Kunda (1990), un ensemble de travaux s'est retrouvé sous l'étiquette du 'raisonnement motivé'. Il s'agit d'une vision du raisonnement proche de celle développée ici, comme en témoigne ce passage dans la conclusion de l'article :

> I have proposed that when one wants to draw a particular conclusion, one feels obligated to construct a justification for that conclusion that would be

> plausible to a dispassionate observer. In doing so, one accesses only a biased subset of the relevant beliefs and rules. (p.493).

Je vais maintenant passer en revue certains des travaux cités par Kunda, ainsi que des études postérieures dans la même tradition, en soulignant au besoin les différences entre l'interprétation qui en est généralement faite et celle qu'il est possible d'en faire dans le cadre de la théorie argumentative. Les expériences qui vont être exposées permettront d'illustrer les différents aspects qui sont importants à la fois pour la théorie de Kunda et pour la théorie argumentative. Il s'agit de montrer tout d'abord qu'il s'agit bien de raisonnement. Ensuite qu'il s'agit de raisonnement *motivé*, et non 'objectif'. On peut ensuite éliminer des explications alternatives. Et enfin, le rôle important joué par la disponibilité des justifications sera souligné.

Une première expérience peut déjà illustrer plusieurs de ces points. Il s'agit de la façon dont sont traités les résultats de tests d'intelligence (Wyer & Frey, 1983). Les participants étaient regroupés en paires acteur/observateur. Les deux membres de chaque paire commençaient par estimer leur propre intelligence, puis recevaient un test d'intelligence. Ensuite, l'acteur recevait un feedback positif ou négatif (selon les conditions). L'observateur avait accès au feedback concernant l'acteur, mais pas au sien. Après cela, les participants devaient lire un rapport contenant des arguments positifs et négatifs vis-à-vis des tests d'intelligence. Enfin, après un court délai, les participants étaient confrontés à une tâche de rappel (portant sur les arguments présentés dans le rapport) assortie de questions sur les tests d'intelligence en général et sur ce test en particulier.

Comme prévu, les participants ayant reçu un feedback négatif, par rapport à ceux ayant reçu un feedback positif, jugèrent les tests d'intelligence en général, aussi bien que le test spécifique auquel ils avaient été confrontés, comme étant moins fiables et moins révélateurs de l'intelligence générale. Les observateurs, par contre, ne firent pas de différence entre les deux conditions, montrant qu'il s'agit bien d'un exemple de raisonnement motivé. De plus, la différence était principalement due à une forte dépréciation des tests en cas de feedback négatif (plutôt qu'à une sur-appréciation en cas de feedback positif), ce qui indique que l'effet vient principalement d'un rejet des résultats dommageable pour l'image de soi (dans ce cas, l'échec relatif à un test d'intelligence). Le résultat des tests de rappel renforce

cette conclusion. Comparant le nombre d'arguments rappelés par les observateurs et les acteurs, on s'aperçoit que lorsque le feedback était négatif, les acteurs rappelèrent plus d'arguments positifs, alors qu'ils en rappelaient moins lorsque le feedback était positif. Les auteurs interprètent ceci comme la marque d'une plus grande élaboration : lorsque le feedback était négatif, les participants auraient cherché à trouver des contre-arguments contre les arguments positifs, facilitant ainsi leur mémorisation (voir Eagly, Chen, Chaiken, & Shaw-Barnes, 1999).

On peut déjà voir dans cette expérience presque toutes les idées qui ont étés mentionnées ci-dessus. Les résultats du test de rappel donnent de bonnes raisons de penser qu'il s'agit bien là de raisonnement et pas de phénomènes de plus bas niveau. En effet, si certains participants avaient uniquement rejeté les arguments en faveur des tests d'intelligence sans y prêter attention, ils devraient moins bien s'en souvenir. Le fait qu'ils s'en souviennent mieux, au contraire, montre qu'ils ont passé du temps à y réfléchir. Or, l'effet de cette réflexion est contraire au sens des arguments. On voit mal comment une telle inversion aurait pu avoir lieu si les participants n'avaient pas cherché, et trouvé, des contre-arguments leur permettant de réfuter les arguments présentés. Par ailleurs, il s'agit bien de raisonnement motivé, car seuls les acteurs ayant reçu un feedback négatif furent touchés, et non les observateurs ou les acteurs ayant reçu un feedback positif. Il ne peut donc s'agir d'un effet des croyances préalables par exemple : les participants étant répartis aléatoirement entre les conditions, il n'y a pas de raisons de penser que leurs croyances préalables vis-à-vis des tests d'intelligence diffèrent.

Une autre étude peut illustrer le rôle crucial joué par la disponibilité des arguments. Il s'agit cette fois d'échecs ou de gains dans des paris sportifs (Gilovich, 1983). Des participants, étudiants mais ayant une très bonne connaissance du football américain, avaient pour première tâche de parier sur divers matchs de la NFL (la ligue de football nord-américaine). Une semaine après avoir fait les paris, et après que tous les matchs se soient joués, les participants durent expliquer le résultat des matchs : dire pourquoi les équipes en compétition avaient gagné ou perdu. Enfin, après une nouvelle semaine, c'est la mémoire des participants qui fut testée par une tâche de rappel.

L'analyse des explications fournit plusieurs résultats intéressants. Tout d'abord, les participants parlaient plus des paris perdus que des paris gagnés (et

l'analyse de la tâche de rappel confirme la plus grande attention portée aux pertes). Et surtout ils n'en parlaient pas du tout de la même façon. Dans le cas des pertes, les participants faisaient surtout des commentaires sur pourquoi les choses auraient dû se passer différemment. A l'inverse, dans le cas des gains, les participants expliquaient pourquoi les choses n'auraient pas pu se passer différemment, ou même que le résultat aurait dû être encore plus extrême dans la direction qu'ils avaient prédite.

Les deux expériences suivantes se concentrent sur l'effet de la disponibilité de certaines justifications sur les explications. En particulier, l'auteur considère l'influence qu'un événement chanceux ('a fluke') peut avoir. Par exemple, dans un des matchs examinés (de basket-ball cette fois), un joueur avait raté un panier à cause de ce qui aurait pu être considéré comme étant une faute mais qui n'avait pas été sifflé. Cet événement arrivant peu avant la fin, on peut considérer que s'il avait été sifflé, le résultat du match aurait pu être différent. L'auteur a profité de ces circonstances pour appeler des personnes ayant vu le match, en leur rappelant (ou non) cet épisode. Il leur demanda si le résultat changerait si le match devait être rejoué. Parmi les personnes soutenant l'équipe l'ayant emporté, le fait de rappeler cet épisode n'eut quasiment aucun effet : de 100% à penser que le résultat serait le même, ils furent encore 90% lorsque l'épisode était rappelé. Par contre, pour les supporteurs de l'équipe adverse, celle qui aurait pu l'emporter si la décision de l'arbitre avait été différente, le fait de rappeler l'épisode permit à près de 50% de personnes de changer d'avis : alors que 70% admettait que leur équipe perdrait si le match était rejoué dans la condition contrôle, ils ne furent plus que 23% lorsque l'épisode était rappelé, leur fournissant un moyen aisé de penser (et de justifier) que le résultat aurait pu être différent.

Les participants dont l'équipe favorite avait perdu étaient tous motivés pour penser qu'elle aurait très bien pu gagner. Cependant, lorsqu'aucune justification n'était facilement accessible, ils ne furent qu'une minorité à donner cette réponse. Par contre, dès qu'une justification leur était fournie, plus des trois-quarts saisirent l'opportunité et donnèrent alors la réponse souhaitée.

On pourrait citer ici de nombreux autres résultats. Markus et Kunda (1986) ont montré que des participants amenés à se sentir trop normaux, ou trop uniques, raisonnaient pour retrouver une vision d'eux-mêmes plus équilibrée (voir aussi Kunda, Fong, Sanitioso, & Reber, 1993). Dans une série d'expériences, Kunda

(1987) montre que les théories causales sont influencées par les motivations des participants, et ce au-delà de tout effet des croyances préalables. Sanitioso, Kunda et Fong (1990) et Ross, McFarland et Fletcher (1981) se sont penchés sur les usages qui pouvaient être faits de souvenirs spécifiques comme arguments permettant de soutenir une vision positive de soi. Ginossar et Trope (1987) ont quant à eux étudié l'utilisation d'indices statistiques comme arguments, utilisation guidée par le raisonnement motivé. De la même façon, Boiney, Kennedy et Nye (1997) ont démontré le rôle du raisonnement motivé dans la prise de décision de nature économique. Dunning, Meyerowitz et Holzberg (1989) ont montré que le raisonnement motivé peut expliquer pourquoi les gens pensent être meilleurs que la moyenne dans certains cas, mais pas dans d'autres. Et pour finir, on peut décrire dans un peu plus de détails une étude montrant que les scientifiques sont, eux aussi, touchés par ce phénomène.

Dans cette étude, d'éventuels biais dans la façon dont les articles sont relus ont été examinés (Mahoney, 1977). L'auteur a utilisé une controverse qui faisait rage à l'époque dans une psychologie qui échappait peu à peu au béhaviorisme. Des résultats paraissaient montrant que, dans certaines circonstances, des récompenses pouvaient avoir l'effet de diminuer la motivation intrinsèque. Cet effet est en flagrante opposition avec les bases mêmes du béhaviorisme, pour lequel le renforcement doit augmenter la motivation. Les participants étaient des relecteurs invités du « Journal of Applied Behavioral Analysis », une revue bien ancrée dans le courant behavioriste. Parmi les cinq conditions mises en place, les deux premières sont les plus pertinentes. Les relecteurs reçurent un article contenant une introduction, une section de méthodologie et les résultats, une excuse ayant été trouvée pour ne pas intégrer de discussion. L'expérience décrite visait précisément la question de l'effet des récompenses sur la motivation intrinsèque chez des enfants. Les deux premières parties étaient identiques, et équilibrées (l'introduction mentionnant autant de travaux allant dans la direction béhavioriste que s'y opposant). Par contre, dans une condition les résultats étaient en faveur de la perspective béhavioriste (résultats positifs : la motivation interne des enfants augmentait avec la récompense), alors qu'ils y étaient opposés dans une autre (résultats négatifs : la motivation interne des enfants diminuait avec la récompense).

Etant donné que les parties introductive et méthodologique étaient identiques, les relecteurs auraient dû attribuer des notes similaires aux articles dans les deux conditions. Si, au contraire, ils raisonnent de façon biaisée, ils devraient trouver plus facilement des arguments soutenant les articles dont les résultats leurs conviennent, ou des arguments attaquant ceux dont les résultats ne leurs conviennent pas (ou faire les deux), mais en tout cas ils devraient attribuer des notes plus basses aux articles dont les résultats les dérangent. Les résultats furent sans appel : dans presque toutes les sections (méthode, présentation des données, contribution scientifique et jugement global), les notes attribuées dans la condition négative étaient significativement inférieures à celles de la condition positive (la seule catégorie non significativement différente étant la pertinence scientifique de l'article). A la méthodologie par exemple, pourtant strictement identique entre les deux versions, ne fut attribuée qu'une seule note de 0 (sur une échelle entre 0 et 6, et parmi 12 relecteurs) dans la condition positive, alors que près de la moitié des membres de la condition négative (6 sur 14) lui attribuèrent cette note (moyennes : 4,2 et 2,4 respectivement). Ces zéros se retrouvent à l'identique pour la note globale, qui est du coup elle aussi très différente selon les conditions (3,2 versus 1,8). Cela signifie que l'article aurait été accepté par la majorité des relecteurs lorsqu'il avait des résultats leur convenant, mais rejeté par une majorité avec des résultats dérangeants.

On peut alors se demander s'il s'agit plus d'une tendance à chercher des défauts lorsque les résultats ne sont pas en accord avec des positions précédentes, ou s'il s'agit au contraire d'un renforcement des arguments qui s'y accordent. Mahoney mentionne une erreur involontaire dans la façon dont les articles ont été rédigés qui permet de répondre à cette question. Dans la description de l'expérience, il est mentionné que huit participants furent utilisés au total. Il s'agit en fait d'une typo, qu'il est possible de repérer car d'une part il est mentionné que les participants sont divisés en trois groupes égaux (difficile avec 8 participants), et d'autre part le tableau de résultats mentionne 12 participants. Si les relecteurs confrontés à des résultats qui ne leur conviennent pas sont plus attentifs, ils devraient être plus nombreux à avoir détecté cette incohérence, ce qui fut le cas en effet : ils furent plus de 70% à la détecter, contre un quart dans la condition positive. Il me semble difficile d'expliquer ces résultats autrement que par une motivation à trouver des arguments pour rejeter les conclusions qui nous déplaisent, motivation qui manque lorsqu'au contraire ces

conclusions nous conviennent (voir également Koehler, 1993, pour une expérience très proche de celle-ci aux résultats similaires).

De toutes ces études, on peut retenir deux éléments importants. Le premier est que les gens utilisent parfois leurs capacités de raisonnement à des fins autres qu'une visée épistémique 'normale'. Ils l'utilisent pour montrer qu'ils ont certaines caractéristiques – soit qu'elles soient naturellement jugées comme désirables, soit que le contexte les ait amenées à être jugées comme telles – ou que les caractéristiques qu'ils ont sont désirables. Ils l'utilisent pour se convaincre que des conclusions indésirables peuvent être rejetées. Ils l'utilisent pour (se) persuader que leurs erreurs n'en étaient pas vraiment. Pour servir ces différents objectifs, ils peuvent par exemple chercher en mémoire des éléments qui pourront leur servir d'argument, ou se montrer très attentifs aux défauts de certains arguments – mais pas d'autres. Le second point important est que cette recherche d'arguments va jouer un rôle dans la réponse des participants : ils n'adoptent pas simplement les croyances qui pourraient les arranger, ils sont contraints par la disponibilité d'arguments pouvant les soutenir (ils doivent maintenir une « illusion d'objectivité », Pyszczynski & Greenberg, 1987). C'est cela qui indique bien qu'il s'agit de raisonnement : les participants n'adoptent pas simplement les croyances désirées, ils ne le font que s'ils trouvent des arguments adéquats. Comme le dit Kunda :

> As will become clear from the work reviewed in this section, an explanation for how directional goals affect reasoning has to account not only for the existence of motivated biases but also for the findings suggesting that such biases are not unconstrained: People do not seem to be at liberty to conclude whatever they want to conclude merely because they want to. Rather, I propose that people motivated to arrive at a particular conclusion attempt to be rational and to construct a justification of their desired conclusion that would persuade a dispassionate observer. They draw the desired conclusion only if they can muster up the evidence necessary to support it. (pp.482-3)

On pourrait penser qu'il y a là matière à débat sur le rôle causal que peut jouer la conclusion. Pour la théorie présentée ici, la conclusion vient en premier, et les arguments la soutenant ne viennent qu'ensuite. Ainsi qu'elle est formulée ci-dessus,

on pourrait penser que pour la théorie du raisonnement motivé les gens ne parviennent à la conclusion qu'après avoir trouvé des arguments. Ce n'est en fait pas le cas : comme le dit Kunda, il s'agit là de personnes *motivées pour parvenir à une certaine conclusion*. Il faut donc bien que la conclusion préexiste à toute recherche d'arguments.

Ces deux caractéristiques pointent vers un mécanisme qui a une fin sociale. On voit mal pourquoi les individus s'engageraient dans cette recherche d'arguments à des fins purement individuelles. D'un côté, si leur objectif était simplement de former des croyances vraies, tous ces raisonnements motivés ne devraient pas du tout exister. D'un autre côté, s'il s'agissait uniquement d'avoir certaines croyances, les gens pourraient simplement modifier leurs croyances, sans s'engager dans ce processus de justification. On pourrait envisager que ce processus de justification soit utilisé pour garantir que les croyances qui sont ainsi formées ne soient pas trop éloignées de la réalité. Mais on pourrait alors se demander à quoi servirait ce semblant d'objectivité ? Pourquoi chercher à avoir des croyances fausses, mais pas trop ? Dans ce cas, à part le maintien d'une certaine plausibilité vis-à-vis d'autres personnes (ce qui est l'hypothèse défendue ici), on ne voit guère de raisons de le faire.

7.2 Effets de l'élasticité des justifications

Dans une série d'articles, Christopher Hsee a étudié l'effet de ce qu'il nomme 'l'élasticité des justifications' (Hsee, 1995, 1996a; Schweitzer & Hsee, 2002). L'idée générale découle de celle du raisonnement motivé : les gens sont souvent motivés pour parvenir à certaines conclusions, pour prendre certaines décisions. Cependant, ils éprouvent le besoin de justifier ces décisions, et hésitent à les prendre s'ils ne peuvent les défendre adéquatement. Le fait qu'un facteur se prête plus ou moins bien à une telle justification (son 'élasticité') devrait donc influer sur la décision. Si une personne a le choix entre A et B, et préfère B, mais que de bons arguments existent pour choisir A, alors cette personne devrait pouvoir tout de même choisir B si un autre facteur lui permet de justifier son choix malgré tout. Donc, la disponibilité, qui dans ce cas est déterminée par l'élasticité, de justifications peut modifier les choix des participants.

Avant de passer aux travaux de Hsee lui-même, on peut mentionner une expérience plus ancienne de psychologie sociale, rapportée dans Hsee (1996a). Dans cette expérience (M. Snyder, Kleck, Strenta, & Mentzer, 1979), les participants arrivaient dans une salle dans laquelle il y avait deux compartiments comprenant chacun une télévision. Dans un compartiment était assis un premier complice, et dans l'autre un second complice portant un appareillage aux jambes marquant un handicap. On disait aux participants qu'ils allaient devoir évaluer des films muets présentés sur les télévisions. Ils devaient d'abord remplir un questionnaire. Pour cela, ils pouvaient s'asseoir à une table triangulaire dont un côté était plus proche du complice prétendument handicapé et l'autre de l'autre. L'hypothèse de base est que les participants ont tendance à s'asseoir près du complice non handicapé. Cependant, les participants ne voudraient pas révéler cette discrimination, et seraient donc plus à même de s'engager dans ce comportement s'ils avaient une autre raison pour le faire (une excuse, une justification potentielle). Ainsi, les expérimentateurs ont comparé deux conditions, une dans laquelle les participants étaient informés que les films montrés sur les deux télévisions seraient identiques, et l'autre dans laquelle on leur décrivait deux films différents (ceux passant sur la télévision du complice handicapé et de l'autre étant contrebalancés). Les résultats confirmèrent les hypothèses : alors que les participants choisissaient les deux côtés équitablement lorsqu'ils n'avaient pas d'excuse leur permettant de révéler leur comportement favori, ils furent significativement plus nombreux à s'asseoir du côté du complice non handicapé lorsque les films étaient différents, et qu'ils avaient donc une raison autre que le handicap d'un des complices pour justifier leur comportement.

Dans une première série d'expériences, Hsee (1995) a étudié l'influence de l'incertitude comme facteur d'élasticité. On peut se contenter de décrire en détail la première expérience. Les participants devaient remplir une tâche assez ennuyeuse de correction d'épreuves : chercher les fautes présentes dans des petites annonces. Les petites annonces concernaient des meubles ou des personnes cherchant un ou une partenaire, les secondes étant faites spécialement pour être considérablement plus amusantes à lire que les premières. Dans une première phase, les participants devaient corriger les textes de deux annonces de chaque type, afin qu'ils réalisent que travailler sur les secondes serait plus intéressant que sur les premières. Ensuite,

ils devaient corriger 10 pages au total de ces deux types d'annonces, pages dont ils pouvaient répartir la quantité à leur convenance entre les types d'annonces. Cependant, avant qu'ils ne fassent leur choix, un écran leur indiquait le nombre moyen d'erreurs par page. Or les participants devaient être payés en fonction du nombre d'erreurs détectées : ils avaient donc avantage à prendre le maximum de pages des annonces contenant le plus d'erreurs. Dans la condition 'fixe', le nombre d'erreur par page ne variait pas et avantageait les annonces de meubles avec 9 erreurs par pages contre 7 pour les annonces personnelles. Dans la condition 'élasticité positive'[53], les quantités étaient les suivantes : annonces de meubles, 9 erreurs par page ; annonces personnelles, entre 1 et 13 erreurs par page. Enfin, pour la condition 'élasticité négative' : annonces de meubles, entre 3 et 15 erreurs par page ; annonces personnelles, 7 erreurs par page. On voit que les quantités moyennes sont identiques dans tous les cas, mais que leur degré d'incertitude est différent. Dans les deux cas élastiques, il est possible de justifier le choix favori par ailleurs (les annonces personnelles car elles sont plus amusantes) en arguant de l'incertitude, et de la possibilité que le nombre d'erreur y soit également supérieur (bien qu'il ne s'agisse pas en fait d'un très bon argument, il a au moins le mérite d'exister).

La prédiction est que les participants choisiront de corriger plus de pages des annonces personnelles lorsqu'ils pensent pouvoir justifier ce choix sur tous les attributs (argent et plaisir) que lorsque l'attribut 'nombre d'erreurs' est totalement inutilisable en ce sens et pointe au contraire dans l'autre direction. C'est bien ce qu'indiquèrent les résultats : alors que les participants de la condition 'fixe' choisissaient de lire en moyenne 2,41 pages d'annonces personnelles, ceux des conditions 'élastiques' en choisissaient 4,30 et 5,19 respectivement (différences significatives entre les deux conditions 'élastiques' et la condition 'fixe').

Dans une seconde expérience, Hsee confirma se résultat en faisant varier un autre facteur d'élasticité et en observant des résultats similaires.

Par la suite, Hsee et ses collègues conduisirent une série d'expériences pour tester la robustesse de ce phénomène d'élasticité des justifications. Hsee (1996a) a ainsi montré que les participants se servent de façon flexible d'attributs positifs et

[53] Ainsi nommée car elle permet au choix étant supérieur sur l'attribut non-pertinent de devenir supérieur à l'autre alternative.

négatifs pour justifier les décisions qu'ils sont motivés pour prendre. En l'absence de flexibilité, ils prennent des décisions moins biaisées. De même, on retrouve des phénomènes similaires dans la façon dont les participants communiquent des informations : ils sont prêts à mentir, ou à présenter des informations de façon biaisée, mais ils le font plus lorsqu'ils pensent pouvoir se justifier (Schweitzer & Hsee, 2002).

Ces études se situent bien dans la lignée des recherches sur le raisonnement motivé. Le fait que la motivation influence directement le comportement des participants n'est pas le point intéressant, mais il est par contre remarquable de constater que cette influence est modérée par la possibilité (ou l'impossibilité) qu'ont les participants de justifier leurs décisions. Elles confirment les résultats mentionnés précédemment montrant que pour que les gens puissent prendre une décision 'biaisée' qui les avantage, il faut qu'ils puissent la justifier. Il est intéressant de constater que ceci est vrai même dans des circonstances dans lesquelles le comportement des participants est totalement individuel. Il est toujours possible que la situation expérimentale introduise l'idée que les résultats vont être observés par quelqu'un, mais dans la majorité des expériences qui viennent d'être décrites (l'exception principale étant celle de M. Snyder et al., 1979), les réponses sont tout à fait anonymes, et on ne demande pas de raisons ou de justifications, ces contraintes sont donc réduites au minimum. Dans ces cas, on peut penser que le raisonnement est également déclenché par un conflit entre des intuitions différentes. D'un côté, les participants comprennent bien quelle réponse leur serait la plus favorable, et d'un autre ils comprennent également quelle réponse pourrait être attendue d'eux. Pour certaines manipulations, il s'agit d'attentes de compétence. Par exemple, dans les expériences de Hsee (1995), les participants comprennent qu'un choix donné maximiserait leurs gains. Il y a alors un conflit entre le fait d'être perçu comme quelqu'un qui maximise ses gains, qui a bien compris comment le faire, qui est compétent, et le fait de choisir la solution qui, intuitivement, est la plus attirante. On pourrait cependant imaginer que le problème soit entre deux intuitions directement : d'un côté la perspective d'une tâche plus intéressante, et de l'autre la perspective de gain. Le raisonnement serait alors utilisé non pour chercher des justifications, mais bien pour chercher à résoudre ce conflit de façon optimale. Mais étant donné que la perspective de gain moyen ne change pas à travers les conditions, on devrait alors

s'attendre à ce que les participants ne modifient pas leur décision en fonction de paramètres non pertinents. Par exemple, dans la première expérience de Hsee (1995), le fait qu'une option contienne de 1 à 13 erreurs par page plutôt que 7 pourrait aussi bien être utilisé pour la choisir que pour la rejeter. Or les participants n'utilisent cette incertitude comme justification que pour l'accepter – car il s'agit dans ce cas de la direction favorisée par leur intuition. Ceci montre bien qu'il y a un conflit entre une alternative qu'ils favorisent intuitivement et une alternative qu'ils pensent être un signe de compétence.

Dans d'autres cas, il s'agit d'un conflit entre une intuition et le fait qu'une décision puisse être prise comme un signe de malveillance ou de malhonnêteté (dans Hsee, 1996a; Schweitzer & Hsee, 2002). A nouveau, la disponibilité de justifications influence le comportement des participants, alors même que les 'données brutes' du problème sont similaires à travers les conditions (elles ne varient pas sur les dimensions qui devraient être pertinentes). Ce résultat montre qu'il ne s'agit pas simplement d'un conflit entre intuitions sur ce qu'il faut faire, mais bien d'un conflit entre des intuitions sur ce qu'il faut faire et des intuitions sur ce que les autres attendent de nous. Le raisonnement est alors activé non pour prendre une décision plus 'rationnelle' (et encore moins plus juste) mais bien pour évaluer la possibilité de justifier une décision que l'on veut prendre dès le début.

Finalement, on peut contraster les utilisations du raisonnement dans ces circonstances et dans celles étudiées dans le cadre du choix basé sur des raisons (voir section 8.2). Dans le cas du choix basé sur les raisons, les participants n'ont généralement pas d'intuitions fortes. Le raisonnement a alors pour effet de favoriser l'option qui est la plus facilement justifiable : selon la disponibilité des justifications, il pourra faire pencher la balance d'un côté ou de l'autre. Le cas des justifications élastiques est différent. Dans ce cas, les participants ne cherchent pas à départager deux intuitions conflictuelles, ils cherchent à trouver des raisons pouvant justifier la prise d'une décision qu'ils favorisent malgré le fait qu'elle ne corresponde pas à des critères de compétence ou de bienveillance. En effet, les justifications rendues possibles dans les conditions 'élastique' n'ont pas de direction intrinsèque. Par exemple, dans la première expérience de Hsee (1995), le fait qu'une option contienne de 1 à 13 erreurs par page plutôt que 7 pourrait aussi bien être utilisé pour la choisir que pour la rejeter. Or les participants n'utilisent ceci comme justification que pour l'accepter – car il s'agit dans ce cas de la direction favorisée par leur intuition.

7.3 Théorie de la quantité de traitement

Plusieurs des études passées en revue jusqu'à présent peuvent être interprétées comme montrant que les gens analysent plus attentivement les informations qui sont contraires à leurs croyances ou, plus généralement, qui sont déplaisantes. Pour certains auteurs, cette simple différence d'activation expliquerait les différences entre la façon dont sont traitées les informations contraires à nos croyances préalables et celles qui au contraire s'y accordent bien (Ditto, Scepansky, Munro, Apanovitch, & Lockhart, 1998). Cela signifie que le fonctionnement même du raisonnement ne serait pas biaisé, mais seulement son activation, ce qui va à l'encontre des prédictions de la théorie argumentative. Ils s'appuient sur une série de travaux (qui ne sera pas examinée ici) qui a montré que les gens ont tendance à s'engager plus facilement dans des analyses cognitives détaillées et demandant des efforts lorsqu'ils sont l'emprise d'une émotion négative – ou simplement de mauvaise humeur (voir N. Schwarz, 1991). Deux articles de Ditto et ses collègues visent à tester précisément l'idée que les gens s'engagent dans des traitements plus approfondis vis-à-vis des conclusions qui s'opposent à leurs croyances ou à leurs préférences. Ces expériences vont être décrites en détail, car j'essaierai ensuite de montrer qu'on peut en tirer d'autres conclusions.

Dans une première série d'expériences, Ditto et Lopez (1992) ont examiné une conséquence de l'hypothèse précédente. Si les gens sont plus sceptiques vis-à-vis des conclusions s'opposant à leurs croyances ou à leurs préférences, ils devraient requérir plus d'informations pour parvenir à une telle conclusion qu'à une conclusion qui au contraire s'accorde bien avec leurs croyances ou leurs préférences. Dans la première expérience, les participants devaient tout d'abord remplir un test de 18 questions analogiques (liées à l'intelligence générale), puis on leur disait qu'ils allaient devoir évaluer deux personnes en s'imaginant qu'ils étaient à la place de l'administrateur chargé de décider quelles personnes accepter à l'université et dans ce cas, plus précisément, celle de deux personnes qui était la plus intelligente. Ils avaient plusieurs éléments afin d'évaluer ces deux personnes : une photo, leurs notes de lycée (GPA), leurs performances à un test de questions analogiques similaire à celui passé par les participants ainsi qu'une feuille décrivant la façon dont ils avaient

été perçus par une personne ayant dû remplir une tâche de résolution de problème en collaboration avec eux.

Les deux éléments cruciaux sont les deux derniers. En effet, il était dit aux participants qu'après avoir choisi une des deux personnes comme étant la plus intelligente, ils allaient devoir affronter une tâche de résolution de problème en coopérant avec cette personne. Il s'agit de la manipulation de motivation cruciale. Dans un cas, les deux personnes à évaluer avaient été décrites par leurs partenaires précédents comme étant également sympathiques, agréables en collaboration. Dans l'autre cas, une personne était décrite en ses termes alors que l'autre était décrite comme étant très désagréable, absolument pas sympathique et se conduisant comme un « je sais tout ». Etant donné que les participants pensaient devoir interagir avec la personne qu'ils choisiraient, ils devraient alors être motivés pour choisir la personne agréable comme étant la plus intelligente. La quantité d'information requise par les participants fut mesurée au moyen de la présentation des résultats des deux personnes aux questions du test analogique. Sous prétexte de simuler les pressions pesant sur les administrateurs chargés de prendre ces décisions, les résultats étaient montrés question par question, et les participants devaient s'arrêter dès qu'ils pensaient avoir suffisamment d'informations pour prendre leur décision. Dans une condition (antipathique-positif), la personne antipathique avait un résultat positif (15 réponses sur 18 correctes) alors que la personne sympathique avait un résultat plutôt négatif (9 réponses sur 18 correctes), et dans une autre (antipathique-négative), les résultats étaient inversés. Enfin, il faut souligner que les autres éléments à la disposition des participants concernant l'intelligence favorisaient très légèrement la personne antipathique.

La mesure pertinente est donc le nombre de questions qui fut nécessaire pour que les participants déterminent quelle personne était la plus intelligente. La comparaison la plus intéressante est celle entre le nombre de questions examinées pour accepter que la personne antipathique mais ayant le plus de réponses correctes est en effet la plus intelligente, et le même nombre mais lorsqu'il s'agissait de la personne sympathique. Alors que les participants requirent plus de neuf questions en moyenne pour accepter que la personne antipathique était la plus intelligente, moins de sept leur suffirent lorsqu'il s'agissait de la personne sympathique. Ce résultat est donc conforme à la prédiction des auteurs que les participants auraient besoin de moins d'information pour parvenir à une conclusion désirable (la personne

sympathique est la plus intelligente) qu'à une conclusion indésirable (la personne antipathique est la plus intelligente).

La deuxième expérience étudie la réaction face à un faux test médical. Les participants étaient informés que des expériences avaient montré un lien entre l'absence d'un enzyme et la survenue de troubles pancréatiques. Ils étaient ensuite amenés à tester dans leur propre salive la présence de cet enzyme au moyen d'une bande de papier trempée dans un échantillon de salive. Dans une condition, un changement de couleur de la bande impliquait la présence de l'enzyme, et dans l'autre son absence. Il s'agissait en fait d'une simple bande de papier qui ne pouvait pas changer de couleur. Après avoir testé la présence de l'enzyme, les participants devaient remplir un questionnaire contenant des items portant sur l'enzyme en question, la dangerosité des maladies pancréatiques, et la validité du test. Les participants amenés à croire que leur salive ne contenait pas l'enzyme se montrèrent plus sceptiques vis-à-vis de la validité du test, et jugèrent les maladies pancréatiques moins dangereuses. De plus, les participants étaient filmés à leur insu alors qu'ils testaient leur salive. Il leur avait été indiqué que si elle devait se colorer, la bande le ferait en moins d'une minute, et que généralement 20 secondes suffisaient. Les participants pour qui une absence de coloration indiquait la présence de l'enzyme (le résultat désiré) attendirent en moyenne 75,5 secondes avant de placer la bande dans l'enveloppe indiquant que le test avait été complété, alors que ceux pour lesquels cette même absence de coloration indiquait l'absence de l'enzyme (un résultat indésirable) attendirent significativement plus longtemps (104,8 secondes). De plus, plus de la moitié (52%) de ces derniers testèrent leur salive une seconde fois, alors qu'ils ne furent que 13% dans la première condition. Il s'agit donc encore d'un cas dans lequel les participants requirent plus d'information pour parvenir à une conclusion indésirable qu'à une conclusion désirable (résultats répliqués dans Ditto, Munro, Apanovitch, Scepansky, & Lockhart, 2003).

La troisième expérience était proche de la seconde. Une première différence vient du fait que dans ce cas l'absence de l'enzyme a des conséquences négatives pour la santé dans une condition et positives dans l'autre. La seconde différence vient des mesures dépendantes : dans cette expérience, les participants devaient indiquer, après avoir fait le test, si des événements récents (dans leur histoire médicale ou leur vie plus généralement) avaient pu affecter le résultat du test. Les participants ayant été amenés à croire que le test était révélateur de conséquences négatives sur la santé

listèrent plus de raisons pour lesquelles le test pourrait ne pas être valide (1,7 en moyenne contre 0,5 en moyenne pour ceux pensant que le test révélait au contraire de possibles conséquences positives). De plus, le nombre de raisons générées était très significativement négativement corrélé avec la validé perçue du test : cela signifie que parmi les participants motivés pour ne pas accepter les conclusions du test, ceux qui purent générer plus de raisons pour lesquelles il pourrait ne pas être révélateur se permirent plus facilement de ne pas prendre en compte son résultat. Ces trois expériences confirment l'idée que les gens requièrent plus d'information pour accepter des conclusions déplaisantes. L'expérience trois est particulièrement pertinente car elle fait explicitement recours à l'utilisation de raisons afin de pouvoir douter d'une conclusion déplaisante, mais on peut également concevoir les deux premières expériences comme des recherches de raisons permettant aux participants de rejeter une conclusion indésirable. Et il est important de souligner que devant leur incapacité de trouver de telles raisons, les participants, dans leur grande majorité, finirent pas accepter la conclusion déplaisante (il ne s'agit donc pas de simple 'wishful thinking', mais bien d'un processus dépendant de l'évaluation de raisons permettant ou non de soutenir une conclusion).

Pour un second article, trois nouvelles expériences furent conduites par Ditto et ses collègues (Ditto et al., 1998). Leur objectif était de départager une hypothèse de traitement biaisé d'une hypothèse (pour laquelle les hypothèses indésirables sont traitées de façon biaisée, avec plus de scepticisme) de différence de quantité de traitement. Pour cette dernière hypothèse, qu'ils défendent, les informations positives ou négatives, plaisantes ou déplaisantes, ne sont pas traitées par des mécanismes différents (le traitement n'est pas biaisé), mais les informations négatives ou déplaisantes sont traitées avec plus d'attention, et cela suffirait à expliquer les asymétries dans la façon dont ces deux types d'informations sont perçus.

Afin de mesurer la quantité de traitement que les participants appliquent à la tâche, la première expérience utilise un paradigme classique d'attribution. Il s'agit simplement d'expliquer le comportement d'une personne. On sait qu'un traitement plus profond permet généralement aux gens de tenir plus en compte des contraintes situationnelles lors du processus d'attribution. L'hypothèse de Ditto et al. est donc que lorsque les participants seront confrontés à un message de valence négative, ils seront plus à même de tenir compte de contraintes situationnelles afin d'attribuer une

attitude à la source du message. Dans l'expérience, les participants devaient commencer par remplir un questionnaire portant sur certaines de leurs attitudes. Ensuite, ce questionnaire devait être examiné par un complice qui donnerait sur cette base un feedback aux participants. En fait, les feedbacks étaient prévus à l'avance de façon à être positifs ou négatifs. De plus, le degré de liberté de l'examinateur était également manipulé : dans une condition, les participants pensaient qu'il avait toute latitude pour former son jugement, alors que dans l'autre les participants étaient informés que l'examinateur avait dû se concentrer sur les aspects positifs ou négatifs du questionnaire. Il était ensuite demandé aux participants de se prononcer sur ce qu'ils pensaient que l'attitude réelle de l'expérimentateur était vis-à-vis d'eux. Si les participants dévouaient plus d'énergie à traiter les informations négatives, ils devraient accorder plus de poids aux contraintes situationnelles lorsque le feedback était négatif, ils devraient donc évaluer l'attitude de l'examinateur comme leur étant moins défavorable lorsque ce dernier était contraint de se concentrer sur les aspects négatifs du questionnaire rempli par les participants. Les résultats furent conformes aux prédictions. Lorsque le feedback était positif, les participants évaluèrent de la même façon l'attitude de l'examinateur vis-à-vis d'eux qu'il ait été contraint de se concentrer sur les aspects positifs du questionnaire ou qu'il ait été totalement libre (absence de prise en compte du contexte, indiquant un traitement superficiel). Par contre, lorsque le feedback était négatif, les participants jugèrent cette même attitude comme leur étant nettement plus défavorable lorsque l'expérimentateur était libre que lorsqu'il était contraint (prise en compte du contexte, indiquant un traitement plus profond).

La deuxième expérience était une réplication de la première en ajoutant un double de chaque condition dans lequel les participants devaient traiter une autre tâche afin de mobiliser leur attention. L'objectif était d'empêcher les participants de traiter le feedback de l'expérimentateur en profondeur et donc de tenir compte des contraintes situationnelles dans le processus d'attribution, ceci afin de montrer, par contraste, que la différence obtenue dans l'expérience précédente est bien due à un traitement plus profond. Les résultats attendus furent obtenus. D'une part les résultats de l'expérience 1 furent répliqués. D'autre part, lorsque les participants avaient une charge cognitive, ils ne prirent jamais en compte les contraintes situationnelles. En particulier, lorsque le feedback était négatif ils attribuèrent la même attitude à l'examinateur qu'il ait été contraint dans ses choix ou non. Ceci confirme bien que la

différence observée dans la première expérience dans l'attribution d'attitude dans le cas du feedback négatif était bien due à un traitement plus profond.

Dans leur troisième expérience, Ditto et ses collègues ont testé la même hypothèse mais d'une façon un peu différente. Cette fois, la différence dans la quantité de traitement devait être révélée non par une différence dans les mécanismes d'attribution, mais par une plus grande prise en compte d'informations statistiques dans une tâche de jugement. Cette expérience était essentiellement identique à celle décrite plus haut (Ditto et Lopez, 1992, expérience 2). Les participants étaient donc confrontés au résultat d'un prétendu test médical (une bande trempée dans leur salive censée révéler la présence ou l'absence d'un enzyme), et ils devaient juger l'efficacité du test. Une des variables était le résultat du test : dans une condition, il prédisait de potentiels problèmes pancréatiques, dans l'autre non. La nouvelle variable introduite dans cette étude est une indication sur la quantité de faux positifs. Dans la condition probable, il est indiqué que le test donnait des résultats faussement positifs une fois sur 10. Dans la condition improbable, il ne s'agissait que d'une fois sur 200. Les auteurs prédisent que lorsque les résultats seront positifs, les participants n'utiliseront pas cette information, alors qu'ils le feront (grâce à un traitement plus poussé) si les résultats sont négatifs. Ce fut en effet le résultat observé : lorsque les résultats étaient positifs, le test était aussi bien évalué quelque soit le taux de faux positifs, alors que lorsque les résultats étaient négatifs, le test donnant le plus de faux positifs était jugé comme étant moins efficace que celui en donnant moins.

La théorie de Ditto et ses collègues est en conflit avec la théorie argumentative, pour des raisons rendues claires par le passage suivant :

> Central to the QOP [Quantity Of Processing] view is an image of people as fundamentally adaptive information processors (Ditto & Lopez, 1992; Lopez, Ditto, & Waghorn, 1994). Whereas past treatments of motivated reasoning have portrayed people as intentionally pursuing the goal of reaching a desired conclusion (i.e., as striving to maintain self-esteem or constructing justifications for preferred conclusions), the QOP view sees the reluctance of people to acknowledge the validity of unwanted information as an unintentional by-product of a quite reasonable strategy of directing detail-

> oriented cognitive processing toward potentially threatening environmental stimuli. (Ditto et al., 1998, p.55)

On voit bien que pour Ditto et ses collègues le raisonnement fait partie d'un ensemble de capacités qui sont adaptatives car elles nous permettent directement de mieux traiter les stimuli en général. Il est essentiellement objectif, et les différences dans le degré d'activation ne font que refléter le fait que certaines situations requièrent davantage de traitement que d'autres. Pour la théorie argumentative, cela peut correspondre à un usage bien spécifique du raisonnement : l'évaluation d'arguments. D'une part, il est logique que le raisonnement soit plus activé lorsque nous examinons des arguments visant à nous convaincre de conclusions opposées à nos croyances ou à nos plans. D'autre part, le raisonnement se doit de conserver alors une part d'objectivité : son rôle est précisément de déterminer quelles conclusions il faut accepter, et d'en accepter certaines qui auraient autrement pu être rejetées car incompatibles avec nos croyances ou nos plans. Cependant, dans les expériences examinées dans le cadre du raisonnement motivé, le raisonnement n'est souvent pas utilisé dans sa fonction d'évaluation, il l'est au contraire dans sa fonction de production. Les participants ne doivent pas évaluer des arguments qui leurs sont présentés, ils utilisent le raisonnement pour donner une réponse justifiable. De plus, même si dans certaines expériences le raisonnement peut-être utilisé en mode d'évaluation, il l'est aussi en mode de production. Or en mode de production il est essentiel qu'il soit biaisé. Plusieurs arguments peuvent être utilisés pour réfuter la théorie de Ditto et ses collègues et montrer que les données soutiennent plutôt la théorie argumentative (de même que les visions plus classiques, comme celle de Kunda, du raisonnement motivé). Je commencerai par avancer des arguments théoriques montrant que la théorie de la quantité de traitement est assez peu plausible a priori. Les données issues de leurs expériences seront ensuite réinterprétées. Enfin, je ferai appel aux résultats d'autres expériences qui contredisent directement la théorie de la quantité de traitement.

Critique de la théorie de la quantité de traitement

Commençons par les arguments les plus généraux. Il semble très peu prudent, dans certains cas tout au moins, de s'attarder à traiter en profondeur des informations déplaisantes. Si vous avez l'impression que quelqu'un va vous attaquer, mieux vaut ne pas y réfléchir à deux fois et réagir immédiatement. A tout le moins, on peut dire que la catégorie 'informations déplaisantes' ne requiert pas tout le temps des traitements plus approfondis et que dans bon nombre de cas, c'est exactement l'opposé qui doit se passer. Ce problème est particulièrement criant pour l'hypothèse mise en avant et testée par Ditto et Lopez (1992) selon laquelle les gens requièrent plus d'information pour parvenir à des conclusions déplaisantes qu'à des conclusions plaisantes. Compte tenu du coût relatif des faux positifs et des faux négatifs, il est évident que dans bon nombre de cas les conclusions déplaisantes seront au contraire atteintes plus facilement. Sur la base d'informations limitées, mieux vaut conclure qu'un champignon comestible est vénéneux que l'inverse. Un domaine dans lequel cela a été prouvé plusieurs fois est celui de la perception des personnes. Comme le notent Fiske et al. (2007) dans leur revue de cette littérature : « Perceivers sensitively heed information that disconfirms, rather than confirms, the other person's warmth ». Dans ce cas, il convient en effet d'être prudent : faire confiance à une personne qui ne le mérite pas peut être très dommageable. Il fait dès lors sens d'accepter plus facilement les conclusions déplaisantes (la personne n'est pas en fait bienveillante) que celles qui sont plaisantes. Notons que dans ce cas le jugement ne se fait pas sur des informations communiquées : il ne s'agit pas de quelqu'un voulant nous convaincre d'une conclusion déplaisante, mais de notre système cognitif en venant lui-même à former une telle conclusion sur la base d'informations perçues.

Pour faire justice aux idées de Ditto et collègues, on peut mentionner un point intéressant qu'ils soulèvent :

> From an adaptive perspective, it would seem counterproductive for an organism to direct its attention to threatening stimuli only to convince itself that the threat did not exist. Rather, the adaptive response to potential threat is to initiate a detail-oriented analysis of it in an attempt to determine whether a threat does indeed exist; if it does, appropriate coping behaviors can be

> initiated, but if it does not, valuable resources will not be expended in mobilizing to cope with an imaginary danger. (Ditto et al., 1998, p.55)

Cette critique s'applique bien à une théorie qui considérerait le raisonnement comme un outil général, visant à traiter tous les types de stimuli rencontrés. Dans ce cas, il est évident en effet que chercher activement à montrer que les informations déplaisantes, ou menaçantes, sont fausses serait terriblement stupide et entraînerait fréquemment une mort précoce. Cependant, pour la théorie argumentative, le raisonnement est surtout fait pour traiter des informations communiquées (qu'il s'agisse de l'évaluation ou de la production). Dans ce cadre limité, un tel biais n'est pas maladaptatif : il est au contraire exactement ce que nous devons attendre du fonctionnement efficace d'un mécanisme cherchant à trouver des arguments pour soutenir une conclusion donnée (ou pour la rejeter). Ce point est important car il souligne le fait que si le raisonnement est bien biaisé vers l'infirmation dans certains cas, il est alors vraiment difficile d'expliquer cela dans le cadre d'une vision classique du raisonnement comme mécanisme visant à traiter des stimuli en général.

Même si la théorie de Ditto et collègues n'est guère plausible dans sa formulation la plus générale, on ne peut pas rejeter aussi facilement les arguments expérimentaux qu'ils avancent pour montrer que le raisonnement n'est pas biaisé. On peut tout d'abord noter qu'il y a une certaine contradiction entre cette position, plus directement défendue dans l'article de 1998, et l'hypothèse qu'ils défendent dans celui de 1992. Rappelons que les expériences de l'article de 1992 démontrent que les participants requièrent plus d'information pour parvenir à des conclusions déplaisantes que plaisantes. Il me semble qu'en fait ces résultats sont difficilement réconciliables avec la théorie de la quantité de traitement. Reprenons leurs expériences une par une. Dans la première, les participants devaient déterminer laquelle de deux personnes était la plus intelligente, et ce alors qu'ils étaient motivés pour en choisir une car elle était plus sympathique. La mesure pertinente était le nombre de réponses à un test d'intelligence qui était requis par les participants avant de faire leur choix. Si on compare ces résultats à ceux d'une condition contrôle (dans laquelle les deux personnes à évaluer étaient également sympathiques), on se rend compte que la seule différence est que les participants requirent moins de questions (moins d'information) lorsque la personne qu'ils étaient motivés pour choisir était en

effet la plus intelligente. Par contre, lorsqu'il leur fallait admettre que la personne qu'ils auraient préféré ne pas choisir était en fait la plus intelligente, ils ne requirent pas plus d'information que dans la condition contrôle. On peut conclure deux choses de ces données. Premièrement, la conclusion déplaisante n'a pas demandé plus de traitement (sous la forme de plus d'informations requises) qu'une conclusion neutre, c'est la conclusion plaisante qui en a demandé moins. Deuxièmement, la différence dans la quantité d'information requise (et donc obtenue) par les participants n'eut aucune influence sur la façon dont ils jugèrent l'intelligence des participants. Si l'information était traitée de façon non biaisée, on pourrait s'attendre à ce que les participants ayant eu accès à moins d'information soient moins sûrs que le participant choisi est en effet plus intelligent que l'autre. En l'absence d'une telle différence, il semble logique de conclure que les participants accordèrent en fait plus de poids aux informations soutenant la conclusion qu'ils voulaient atteindre.

On peut faire une analyse similaire des deuxième et troisième expériences de cet article. Il s'agissait alors de la réaction face à un test censé révéler (ou non) une déficience enzymatique aux effets néfastes (ou bénéfiques pour l'expérience trois). Dans l'expérience trois, les participants pouvaient générer des raisons de refuser les résultats du test, et ils en générèrent plus lorsque ce résultat était de mauvais augure que dans le cas contraire. Cependant, les auteurs avaient également demandé à d'autres participants de générer ces mêmes raisons *avant* de faire le test. Il est donc possible de comparer le nombre de raisons générées par ces participants et par ceux ayant eu à donner des raisons après avoir pris connaissance des résultats du test. Il se trouve que l'évolution la plus importante fut dans le sens d'une diminution lorsque le test était de bon augure (0,5 raisons, alors qu'il y en avait 1,25 en moyenne avant le test) qu'une augmentation lorsque le test était de mauvais augure (1,7 raisons). Là encore, il ne s'agirait donc pas tant d'une augmentation des efforts lorsqu'il s'agit d'éviter une conclusion déplaisante plutôt que d'une diminution lorsque ces efforts pourraient au contraire nous priver d'une conclusion plaisante (dans ce cas en fournissant des raisons de douter des résultats d'un test de bon augure).

Une explication similaire peut être avancée pour les expériences un et deux de Ditto et al. (1998). Dans ces expériences, les participants devaient évaluer l'attitude d'un examinateur vis-à-vis d'eux, examinateur ayant été contraint ou non de leur donner un feedback positif ou négatif. Ditto et ses collègues observent que les participants ne tinrent compte des facteurs contextuels dans l'attribution d'attitude

que lorsque le feedback était négatif. Cependant, pour savoir s'il s'agit d'un écart dans le sens de plus de traitement dans le cas du feedback négatif (conclusion déplaisante) ou de moins de traitement dans le cas du feedback positif (conclusion plaisante), il faudrait avoir un point de comparaison avec une conclusion neutre. Un tel point de comparaison, imparfait, peut nous être fournit par les autres études portant sur l'attribution. Par exemple, l'étude de Gilbert et collègues sur laquelle Ditto et al. s'appuient pour leur manipulation de charge cognitive (Gilbert, Pelham, & Krull, 1988). Dans leur première expérience, les participants étaient confrontés à une tâche d'attribution classique, mais dans une condition ils devaient la remplir avec une charge cognitive importante. Si, conformément aux résultats de Ditto et al., la charge eut l'effet d'annuler toute prise en compte de la situation dans le processus d'attribution, dans la condition contrôle les participants la prenaient bien en compte. Or dans cette tâche la conclusion était neutre pour le participant. Cela signifie que les participants des expériences de Ditto et al. ne traitaient pas particulièrement profondément les informations déplaisantes, mais au contraire qu'ils traitaient plus superficiellement les informations plaisantes. Dans le cas des informations plaisantes (le feedback positif), les participants devaient même être activement motivés pour ne pas prendre en compte les contraintes situationnelles car le faire signifierait diminuer leur propre mérite (le bon feedback ne venant alors pas de leurs propres qualités mais de ces contraintes situationnelles).

Finalement, on peut encore proposer la même analyse pour l'expérience trois de ce même article. Il s'agissait cette fois de la prise en compte de différents taux de faux positifs pour l'évaluation de l'efficacité d'un test. Si le test avait un résultat positif, les participants ne prenaient pas en compte ce taux de faux positifs, alors qu'ils le faisaient en cas de résultats négatifs. Cependant, si cette différence était réellement due à une plus grande quantité de traitement fournie lorsque le test avait un résultat de mauvais augure (conclusion déplaisante), on devrait s'attendre à ce que les participants confrontés à ce résultat, comparés à ceux recevant un résultat rassurant, jugent non seulement le test comme étant moins efficace en cas de taux de faux positifs élevé, mais également à ce qu'ils le jugent comme étant plus efficace dans le cas contraire. Or la seule différence apparaît dans le second cas. L'interprétation la plus naturelle des résultats est dès lors plutôt la suivante : confrontés à une conclusion déplaisante, les participants sautèrent sur l'opportunité de la prendre moins en compte en utilisant le fort taux de faux positifs, mais ils

s'abstinrent bien d'utiliser cette information lorsqu'elle ne faisait que conforter le résultat déplaisant. De plus il est fort possible que l'absence de différence observée lorsque la conclusion était positive ne reflète pas une absence de raisonnement, mais au contraire une utilisation active du raisonnement afin de trouver des raisons de ne pas tenir compte du fort taux de faux positifs et de juger malgré tout le test très positivement. Quoiqu'il en soit, les résultats sont plus révélateurs d'un raisonnement biaisé que d'une simple différence de quantité de traitement.

Tous ces résultats sont donc en fait plus compatibles avec une vision classique du raisonnement motivé, à la Kunda, mais également avec la théorie argumentative, qu'ils ne le sont avec la théorie de la quantité de traitement. Ils confirment cependant un argument déjà fait par les études passées en revue précédemment dans la section sur le raisonnement motivé. Même s'ils sont motivés pour parvenir à une certaine conclusion, les participants ne se permettent pas de donner n'importe quelle réponse : il faut qu'ils puissent justifier leur réponse adéquatement. Dans le cas contraire ils sont contraints de donner une réponse qui les satisfait peut-être moins, mais qu'ils peuvent justifier.

Pour finir cette critique de la théorie de la quantité de traitement, on peut mentionner des études sur le biais d'infirmation en psychologie sociale qui montrent directement que le raisonnement, qu'il soit plus ou moins activé selon les circonstances ou non, est bel et bien biaisé.

Dans une expérience de Cacioppo et Petty (1979), les participants étaient confrontés à un message concernant l'augmentation des frais de scolarité. Un message était contre-attitudinal (augmenter les frais de scolarité pour pourvoir aux besoins de l'université) et un message pro-attitudinal (augmenter les taxes sur des produits de luxes dans le même objectif). Les participants devaient écouter le message persuasif puis lister leurs pensées sur le sujet. Les résultats furent les suivants : les participants listèrent plus de contre-arguments lorsque le message était contre-attitudinal (2,25) que lorsqu'il était pro-attitudinal (1,75)[54]. Cette fois par

[54] Les chiffres présentés ici sont des approximations tirés des graphiques de l'article, les chiffres n'étant pas indiqués. De plus, les auteurs ont également manipulé le nombre de présentations, mais je m'en tiens ici à la condition dans laquelle le message n'a été présenté qu'une seule fois.

contre les auteurs indiquent le nombre total d'arguments, et c'est le message pro-attitudinal qui a évoqué le plus de pensées chez les participants (4,65 au total, contre 4,35 pour la condition contre-attitudinale). Même si cette dernière différence n'est pas significative, ce résultat indique tout de même que l'augmentation du nombre de contre-arguments dans la condition contre-attitudinale ne peut être expliquée simplement par une plus grande activité dans cette condition (voir également Brock, 1967). Une étude plus récente a confirmé ces résultats en se penchant sur la façon dont sont traités et mémorisés les messages pro- et contre-attitudinaux (Eagly, Kulesa, Brannon, Shaw, & Hutson-Comeaux, 2000).

Finalement, une expérience importante d'Edwards et Smith (1996) porte directement sur le point en question ici. Les auteurs avancent un modèle de la façon dont les gens évaluent les arguments qui fait les quatre prédictions suivantes :

> Hypothesis 1: Arguments that are compatible with a person's prior belief will be judged to be stronger than those that are incompatible with a person's prior belief.
> Hypothesis 2: People will take longer to evaluate an argument that is incompatible with their beliefs than an argument that is compatible with their beliefs.
> Hypothesis 3: People will generate more thoughts and arguments when an argument is incompatible with their beliefs than when it is compatible.
> Hypothesis 4: Among the thoughts and arguments generated, more will be refutational (rather than supportive) in nature when the presented argument is incompatible with prior beliefs than when it is compatible. (Ibid., p.7)

Lors d'un pré-test, des participants durent remplir un questionnaire portant sur leurs attitudes vis-à-vis de nombreux sujets. Ce pré-test permit d'établir leurs croyances préalables, ainsi que leur force. Dans la première phase de l'expérience elle-même, les participants devaient évaluer des arguments (présentés sur ordinateur) semblables à celui-ci : « Implementing the death penalty means there is a chance that innocent people will be sentenced to death. Therefore, the death penalty should be abolished ». Sept sujets étaient abordés, avec pour chacun un argument pro et un argument anti (pour ou contre l'abolition de la peine de mort par exemple). Les

participants devaient noter la force de l'argument, et les instructions insistaient bien sur le fait que cette évaluation devait se faire indépendamment de leur jugement sur la vérité de la conclusion. A l'issue de cette première phase, les participants recevaient un livret de sept pages avec en tête de chaque page la conclusion d'un des arguments présentés. Il pouvait donc s'agir soit de la conclusion de l'argument pro soit de celle de l'argument anti. Les participants avaient alors trois minutes pour noter autant de pensées que possible sur le sujet.

Les quatre hypothèses furent confirmées. Les arguments compatibles avec les croyances préalables des participants furent notés comme étant nettement plus convaincants que ceux qui étaient incompatibles : les notes attribuées aux arguments compatibles étaient en moyenne plus de deux fois plus élevées que celles attribuées aux arguments incompatibles[55]. De plus, les participants passèrent plus de temps à lire les arguments incompatibles que compatibles. Concernant la seconde phase de l'expérience (lister les pensées), on observe que les participants générèrent plus de pensées lorsque la conclusion était celle d'un argument incompatible que compatible. De plus, les arguments générés tendaient à soutenir la conclusion lorsqu'elle était compatible et à la réfuter lorsqu'elle ne l'était pas, comme le montre clairement la figure suivante (tirée de Edwards et Smith, 1996, p.12) :

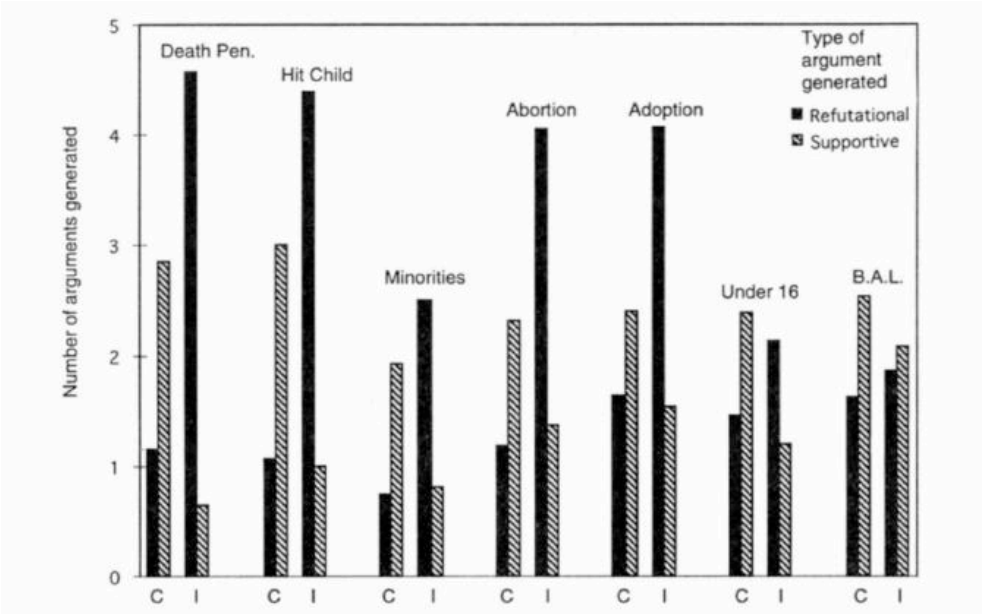


Figure 5. Mean number of supportive and refutational arguments generated in response to compatible (C) and incompatible (I) arguments: Experiment 1. Pen. = penalty; B.A.L. = blood alcohol level.

[55] En excluant de l'analyse deux sujets pour lesquels les participants avaient des croyances préalables peu marquées.

Bien qu'ils soient très concluants vis-à-vis des hypothèses de départ, on peut soulever aux moins deux questions concernant ces résultats. La première concerne la nature du biais dans la recherche d'arguments. Pour les auteurs, les gens cherchent préférentiellement en mémoire des arguments réfutant les conclusions incompatibles avec leurs croyances. On pourrait cependant imaginer une explication inverse. Les sujets utilisés ici étaient très polarisants, la plupart des participants se plaçant franchement pour ou contre. On peut dès lors imaginer qu'une des raisons pour laquelle un participant a une opinion sur un sujet donné est qu'il a des arguments contre la position opposée, ou que, ayant une position donnée, il soit surtout confronté (et donc conserve surtout en mémoire) des arguments contre la position opposée. Dès lors, une recherche d'arguments en mémoire, même non biaisée, pourrait expliquer les résultats obtenus. Si une personne possède en mémoire de nombreux arguments contre l'abolition de la peine de mort, et presque aucun pour, il n'est pas nécessaire de faire appel à un mécanisme de recherche biaisé pour expliquer que cette personne produise spontanément davantage d'arguments contre que pour[56]. Bien que les deux explications ne soient pas incompatibles (en fait, elles se renforcent même mutuellement), la théorie d'Edwards et Smith (de même que la théorie argumentative) prédit l'existence d'un biais dans la recherche, en plus de tout biais dans la répartition des arguments en mémoire.

Heureusement, un résultat obtenu par hasard suggère qu'un biais dans la recherche existe bel est bien. Les auteurs ont en effet observé que les participants généraient parfois des arguments redondants. S'il ne s'agissait pas de doublons exacts ou de simples paraphrases, il s'agissait d'arguments très similaires, dont on peut considérer qu'ils étaient deux instances d'un même argument. De tels arguments indiquent, me semble-t-il, la présence d'un biais motivationnel : les participants qui les donnent sont motivés pour donner le plus d'arguments possible, quitte à se répéter un peu. Or ces arguments furent beaucoup plus nombreux lorsqu'il s'agissait pour les participants de réfuter une conclusion incompatible. De plus, une seconde expérience très proche de celle qui vient d'être décrite montra que la très grande

[56] Un des avantages qu'offre l'étude des syllogismes abstraits est justement d'éviter ce genre de problème, mais elle a d'autres limites. Il est cependant rassurant de voir que les conclusions de l'étude d'arguments abstraits et concrets semble converger, au moins pour ce qui est de la présence et de la direction des biais.

majorité de ces arguments étaient générés lorsqu'il s'agissait d'un argument à la conclusion incompatible *et* très chargé émotionnellement. En l'absence de cette seconde condition, presque aucun argument redondant n'était généré, même dans le cas des arguments aux conclusions incompatibles. Ce résultat est plus cohérent avec l'hypothèse d'une motivation qui biaise la recherche qu'avec l'hypothèse de différences dans le nombre d'arguments mémorisés.

Une autre question qu'on peut se poser vis-à-vis de l'expérience d'Edwards et Smith est la suivante : dans quelle mesure teste-t-elle uniquement l'évaluation d'arguments ? S'il est indéniable qu'une phase initiale d'évaluation est présente lorsque les participants lisent les arguments, il semble très probable qu'ensuite le raisonnement soit utilisé en production, afin de chercher à justifier la réponse initiale formée lors de la phase d'évaluation (qui ne comprend d'ailleurs pas que du raisonnement, mais également des mécanismes d'évaluation comme la vérification de cohérence). Dans ce cas, il est fort possible que les biais observés dans cette expérience résultent en fait davantage de la phase durant laquelle le raisonnement est utilisé en production que de celle durant laquelle il est utilisé en évaluation. Cependant, la théorie de la quantité de traitement de Ditto et ses collègues s'applique au raisonnement en général, et ces résultats conservent donc toute leur pertinence en ce qu'ils s'opposent directement à cette théorie. S'ils montrent bien, conformément aux prédictions de la théorie de Ditto que le raisonnement est plus activé lorsque la conclusion est incompatible (ce qui est cohérent avec la théorie argumentative), ils montrent également qu'il est aussi biaisé dans son fonctionnement, ce qui contredit directement la théorie de la quantité de traitement.

On pourrait rappeler d'autres résultats, parmi ceux précédemment décrits, allant à l'encontre de la théorie de la quantité de traitement. Les résultats de Gilovich (1993), par exemple, montrèrent une claire asymétrie dans la façon dont les participants expliquaient leurs pertes et leurs gains. Dans sa formulation la plus générale, la théorie de la quantité de traitement n'est donc guère plausible, et elle est contredite par plusieurs résultats expérimentaux. Ironiquement, on peut retenir les résultats des expériences de Ditto et collègues comme d'autres soutiens de la théorie du raisonnement motivé et, partant, de la théorie argumentative.

Vont maintenant être décrites une série de conséquences que peut avoir l'utilisation du raisonnement motivé. L'examen de ces études est intéressant pour deux raisons principales. D'une part, elles montrent à quel point les utilisations 'motivées' du raisonnement sont communes et robustes. Il peut expliquer plusieurs biais récurrents en psychologie, comme l'évaluation biaisée d'arguments, la polarisation des attitudes, la persévérance des croyances ou la sur-confiance. D'autre part, elles montrent que le raisonnement a souvent des conséquences épistémiques négatives. Tous ces résultats rendent donc beaucoup plus difficile de s'en tenir à une vision du raisonnement comme d'un instrument visant à améliorer la qualité de la cognition individuelle.

7.4 Conséquences pour l'évaluation et la génération d'arguments

Un des domaines dans lequel le raisonnement motivé (qu'il soit ainsi nommé ou non dans ces travaux) a été le plus étudié est celui de l'évaluation d'arguments. On a déjà noté, dans la partie portant sur les capacités à évaluer et produire des arguments, que les gens étaient souvent biaisés – par exemple les gens ne pensent que rarement spontanément à des contre-arguments contre leurs théories. Les études qui vont être examinées maintenant portent plus précisément sur la façon dont les participants examinent et évaluent des arguments, arguments dont les conclusions peuvent être plus ou moins plaisantes pour les participants – elles peuvent par exemple remettre en cause leur religion, leur choix de carrière, ou au contraire s'y accorder. Au-delà du fait que les participants tendent à évaluer plus négativement les arguments aux conclusions déplaisantes, est plus intéressant le fait que cela les pousse à chercher des défauts dans ces arguments : ils ne se contentent pas de les rejeter, il faut qu'ils trouvent des raisons pour le faire. Le caractère déplaisant de la conclusion aurait un effet motivationnel sur l'évaluation que les participants souhaitent en faire (l'évaluer négativement), mais ils ne se contenteraient pas de la rejeter : ils chercheraient des moyens de justifier ce rejet. C'est pour cette raison que ces travaux s'accordent bien avec les hypothèses de la théorie du raisonnement motivé et, partant, de la théorie argumentative.

A la suite de ces recherches, nous nous intéresserons à d'autres expériences portant davantage sur les arguments qui sont générés par les participants lorsqu'on

leur demande de penser à des arguments pour ou contre une certaine idée. Nous verrons là encore que les participants tendent à donner des arguments qui leur permettraient de justifier leur point de vue initial.

Paul Klaczynski et ses collègues ont conduit une série d'études portant sur l'influence de divers facteurs (âge, intelligence, etc.) sur le développement et l'efficacité des capacités de raisonnement. Un facteur est particulièrement pertinent ici : le rôle que joue la position des participants vis-à-vis de la conclusion des arguments examinés. Ce facteur permettra d'évaluer de possibles biais dans l'évaluation des arguments. Etant donné leurs points communs, tant dans les objectifs que dans les méthodes, je me contenterai de décrire rapidement ces expériences et d'exposer les résultats les plus saillants dans l'ordre chronologique.

Dans la première étude (Klaczynski & Gordon, 1996b), des adolescents (16 ans en moyenne) devaient évaluer des arguments dont la conclusion était positive, neutre, ou négative vis-à-vis de leur religion. Ces arguments décrivaient une (fausse) étude scientifique comparant des personnes appartenant à différentes religions. Trois études étaient ainsi présentées, ayant chacune un défaut différent permettant d'en réfuter les conclusions (dans un cas par exemple les membres des différentes religions n'étaient pas sélectionnés de la même manière, impliquant la possibilité de rendre compte des résultats par un biais de sélection). Les participants devaient noter la force et la validité de ces différents arguments, puis ils devaient expliquer et justifier leur réponse. Ces réponses furent ensuite codées pour déterminer ceux qui étaient effectivement parvenus à détecter les défauts des arguments. L'orientation de la conclusion eut un effet important sur les évaluations de force et de validité, l'effet étant en majorité dû à des notes beaucoup plus basses pour les arguments aux conclusions négatives comparées aux neutres et positives (les neutres n'étant que non-significativement inférieures aux positives). De façon plus intéressante, on retrouve cette même distinction pour ce qui est de la capacité à repérer des défauts : les participants confrontés à des conclusions négatives furent significativement plus capables de repérer les défauts des arguments que ceux des deux autres conditions. Il semble donc que les participants ne se contentent pas d'évaluer les arguments sur la base de leur conclusion, mais plutôt que la conclusion oriente leur raisonnement, et

que c'est (au moins en partie) parce qu'ils trouvent des défauts qu'ils attribuent des notes plus basses aux arguments dont les conclusions leur déplaisent[57].

Les auteurs concluent de ces résultats que les participants raisonnent davantage lorsqu'ils jugent un argument à la conclusion négative que positive. Les conclusions négatives motiveraient les participants à raisonner, et non les conclusions positives. Cependant, pour que cette interprétation soit valide, il faut penser que le raisonnement, une fois déclenché, est impartial : qu'il a autant de chances de trouver des défauts quelque soit la condition. Aucun élément indépendant n'est donné pour soutenir cette hypothèse. Selon la théorie argumentative, au contraire, le raisonnement peut non seulement être plus ou moins déclenché selon les conditions, mais son fonctionnement même est biaisé : il est orienté vers la recherche d'arguments pour la conclusion que nous voulons soutenir. On peut donc parfaitement imaginer que les participants raisonnent autant dans les différentes conditions, mais que le raisonnement cherche des arguments différents selon les cas. Il n'est alors guère surprenant qu'on passe à côté des défauts d'une étude que l'on scrute pour montrer à quel point elle est bien faite (dans le cas ou sa conclusion nous arrange). A l'inverse, il est naturel de trouver des défauts lorsqu'on les cherche. Etant donné que la seule mesure de raisonnement utilisée ici est le fait de repérer ces défauts, les données ne sont pas concluantes.

Dans le cas du biais de croyance dans les syllogismes classiques, les résultats présentés plus haut portant sur le temps de réponse et la diversité dans la représentation des prémisses tendent à accréditer cette hypothèse : les participants raisonnent toujours autant, mais ils ne cherchent pas la même chose. Cependant, dans le cas des syllogismes, on peut imaginer que l'asymétrie dans la motivation des participants à raisonner sur des conclusions crédibles ou non crédibles est limitée : dans tous les cas, ils ne se sentent pas directement concernés, et même s'ils échouent à trouver des arguments pour réfuter une conclusion telle que « tous les poissons sont des truites », il est clair qu'ils ne l'accepteront jamais, qu'elle ne les influencera nullement. Par contre, dans le cas de l'expérience de Klaczynski et Gordon, le sujet est la religion des participants. Un bon argument, le résultat d'une étude bien

[57] Voir Klaczynski et Narasimham (1998) pour une étude similaire, aux résultats similaires, incluant des participants plus jeunes (de 10 et 13 ans), et Klaczynski et Robinson (2000) pour des populations plus âgées (47 et 70 ans) avec encore une fois des résultats similaires.

contrôlée, ne peut peut-être pas être rejeté sans raison : il ne s'agit plus là uniquement d'un jeu logique, des croyances qui peuvent être assez centrales sont menacées. Il est donc raisonnable de penser que, dans ce cas, les participants seront plus motivés pour raisonner dans le cas des conclusions négatives. Pour raisonnable que cette hypothèse soit, il n'en reste pas moins que les données ne permettent pas de trancher entre les deux hypothèses suivantes. D'un côté une différence de motivation initiale qui ferait que les participants raisonneraient plus lorsque les conclusions sont négatives, mais le ferait de la même façon que lorsqu'elles sont positives. De l'autre côté une différence d'orientation du raisonnement qui ferait que, bien que les participants raisonnent autant dans les différentes conditions, ils ne trouvent des défauts que lorsqu'ils les cherchent, c'est-à-dire dans la condition négative. Et il est également tout à fait possible que le raisonnement soit biaisé aux deux niveaux : que plus d'efforts de raisonnement soient consacrés aux conclusions négatives, mais que quelque soit l'effort consacré le raisonnement cherche principalement à trouver des arguments de soutien pour la conclusion souhaitée[58].

A la suite de ces expériences, Klaczynski et ses collègues en ont conduit de nombreuses autres sur un même modèle. Ils ont mené à bien ces études sur des participants d'âges différents (de 15 à 70 ans), avec des résultats concordants quel que soit la tranche d'âge étudiée. Ils ont montré que les mêmes biais s'observaient lorsque les défauts qu'il fallait trouver dans les arguments présentés prenaient la forme de principes de statistiques (Klaczynski & Gordon, 1996a; Klaczynski, Gordon, & Fauth, 1997; Klaczynski & Lavallee, 2005 ; Klaczynski & Robinson, 2000) ou de paralogismes de l'argumentation (Klaczynski, 1997).

A la suite de Klaczynski, Keith Stanovich et plusieurs de ses collègues se sont intéressés aux liens entre biais de raisonnement et diverses mesures d'intelligence, de

[58] Klaczynski et Gordon mentionnent rapidement ce débat (en termes de 'level of processing' versus 'differential path'), mais en défendant leur hypothèse (level of processing) ils ne tiennent pas compte du fait que la façon dont sont codées les réponses biaise très fortement les résultats en sa faveur (mais ils admettent que les résultats ne permettent pas de trancher et que les hypothèses sont compatibles – pp.332-3). Voir les critiques qui ont été adressées plus haut à la théorie de la quantité de traitement pour des arguments forts montrant que si différence de quantité de traitement il y a, elle n'explique pas tout et que la façon de raisonner elle-même est également biaisée.

traits de personnalité ou autres différences interindividuelles. Toplak et Stanovich (2003) ont ainsi étudié les biais dans la génération d'arguments. Ils ont demandé aux participants leur opinion sur trois sujets de débats (l'augmentation des frais pour l'université, la vente d'organe et l'augmentation du prix de l'essence), puis leur ont demandé de lister des arguments pour et contre une position prise sur chacun de ces sujets. Il n'est guère surprenant de constater que les participants donnèrent plus d'arguments soutenant leur position que d'arguments allant dans l'autre sens. Bien qu'il puisse s'agir d'un réel biais dans la recherche d'argument (et il s'agit sûrement en partie d'un tel biais), ce phénomène peut avoir plusieurs autres explications. Il peut s'agir d'un artefact expérimental : on demandait leur opinion aux participants avant de leur demander les arguments, il est donc possible que, par soucis de cohérence, ceux-ci aient voulu donner un plus grand nombre d'arguments soutenant la position qu'ils venaient de prendre. Il serait aussi normal que la causalité joue dans l'autre sens : ce n'est pas seulement parce que les participants ont une opinion donnée qu'ils ont plus d'arguments pour la soutenir, mais également parce qu'ils ont été exposé à, ou on accepté, plus d'arguments pour la soutenir qu'ils ont cette opinion. Il s'agit dans les trois cas de débats souvent mentionnés dans la presse, qui peuvent être sujets de conversation, et on peut donc s'attendre à ce que les participants aient été exposés à des arguments – qu'ils ne les aient pas créés de toutes pièces. Or différents participants ont très bien pu être exposés à des sources ayant des opinions, et donnant des arguments, différents, ce qui pourrait suffire à expliquer l'effet. Il est très vraisemblable que la causalité joue dans les deux sens, mais les données présentes ne permettent pas de trancher.

Cette étude présente également un résultat intéressant, bien qu'il ne permette pas de résoudre la question de la causalité qui vient d'être soulevée. Bien qu'une corrélation fut observée entre le nombre total d'arguments donnés pour les trois problèmes (certaines personnes donnent généralement plus d'arguments que d'autres), il n'y eut aucune corrélation entre la quantité d'arguments positifs et négatifs. Cette absence globale de corrélation s'explique par un équilibre entre deux facteurs. D'une part la tendance générale qu'ont certains individus à trouver plus d'arguments. D'autre part une corrélation négative entre le nombre d'arguments positifs et négatifs trouvés par chaque individu pour chaque argument. Cela s'explique très bien si on pense que le raisonnement est biaisé pour chercher des arguments soutenant nos positions, et qu'il est biaisé dans la mesure de notre position

initiale : plus nous sommes biaisés (plus notre opinion initiale est forte), plus nous trouverons d'arguments positifs et moins spontanément nous trouverons des arguments négatifs. On pourrait cependant également imaginer une explication en termes d'exposition : si par exemple les arguments venaient de journaux, les participants les plus libéraux (ou les plus conservateurs) pourraient être exposés à un échantillon plus biaisé d'arguments.

Parmi de nombreuses expériences menées à bien dans une étude de grande ampleur de Stanovich et West (2008a), en figure une portant sur la « biais partisan » (myside bias). Il est nécessaire de faire ici un aparté pour clarifier l'emploi de ce terme. Baron (1995) fut le premier à produire une étude portant directement sur le biais partisan, et il crédite Perkins (1989) pour l'introduction du terme. Pour eux, il s'agit d'un biais bien spécifique : « A person may think of good arguments (or bad ones) only on one side of an issue » (Baron, 1995, p.3). Perkins avait observé ce biais en demandant à ses participants de donner des arguments à propos de questions controversées. Il avait également noté que si les instructions demandaient spécifiquement de lister des arguments pour et contre, les participants, bien qu'ils soient toujours biaisés, étaient capables de générer des arguments pour l'autre côté – montrant ainsi qu'il ne s'agissait pas uniquement d'une question de capacité, les arguments étant bien disponibles, mais d'une question de motivation. Le terme est cependant parfois employé dans un sens beaucoup plus large, tellement large qu'il recoupe le biais de confirmation. Ainsi, Stanovich et West l'introduisent de cette façon : « People display myside bias when they generate evidence, test hypotheses, and evaluate policies in a manner biased toward their own opinions » (Stanovich & West, 2008b, p.681). Le problème est qu'élargir de cette façon la définition du biais partisan leur permet d'utiliser des problèmes qui ne font pas nécessairement appel à la recherche d'arguments.

Dans cette expérience, ils ont ainsi recours à des problèmes portant sur l'interdiction potentielle par le ministère des transports d'une voiture dont on a montré qu'elle était plus dangereuse que les autres pour les autres conducteurs. Le biais est créé par le fait que dans un cas il s'agit d'une voiture américaine rejetée par les autorités allemandes, et dans l'autre cas d'une voiture allemande rejetée par les autorités américaines (les participants étant américains). Et effet les participants avaient tendance à davantage être en accord avec les autorités américaines qu'avec

les autorités allemandes. Cependant, rien ne dit qu'il s'agit ici d'un biais dans le raisonnement en soi, dans la recherche d'arguments (ce qu'était originellement le biais partisan, et ce qui nous intéresse ici) : il serait tout à fait possible d'obtenir un résultat similaire avec un raisonnement non biaisé mais se basant sur des préjugés négatifs envers les voitures ou les autorités allemandes ou positifs vis-à-vis des voitures ou des autorités américaines. En effet, bien que les données présentes dans les problèmes même (les informations concernant les voitures) soient identiques dans les deux cas, rien n'empêche les participants d'avoir recours à d'autres informations, et ces informations peuvent parfaitement être biaisées (ou même être exactes en fait, s'il s'avérait que certaines autorités étaient plus ou moins fiables).

Le même problème se pose dans deux autre articles (Stanovich & West, 2007, et Stanovich & West, 2008b, expérience 1) dans lequel les auteurs utilisent des propositions qui risquent d'être perçues différemment par différents participants (une assertion sur les bienfaits de la religion pour des participants religieux et non religieux par exemple). On trouve en effet un biais, dans la direction attendue, dans les notes reflétant l'accord des participants avec les énoncés qui sont attribuées à ces différents énoncés. Mais le raisonnement n'est pas nécessairement impliqué : il ne s'agit pas d'examiner des arguments, mais simplement des propositions. Et, à nouveau, ces résultats pourraient très bien entre obtenus avec des mécanismes de raisonnement non biaisés fonctionnant sur la base de préjugés, ou de connaissances qui seraient, elles, biaisées.

Stanovich et West (2008b, expérience 2) ont également étudié un autre biais, en lien avec le biais partisan, mais qui se situe cette fois dans l'évaluation des arguments. Est-ce que les participants préfèrent les arguments présentant des éléments pour les deux côtés, ou est-ce qu'ils préfèrent les arguments dont les éléments vont tous dans la même direction ? Ils s'inspirent en cela d'une méthode déjà utilisée par Baron (1995) et qui consiste à faire évaluer par les participants des ensembles d'arguments présentés comme venant d'un autre participant. Parmi ces ensembles d'arguments, certains ne contiennent que des arguments allant dans un sens, alors que d'autres sont plus (ou même parfaitement) équilibrés. Les résultats de cette étude sont surprenants : plus les ensembles d'arguments présentés sont biaisés, plus ils ne présentent qu'un côté, plus les participants les jugent bons. Ils ne les jugent pas seulement plus persuasifs, ce qu'on pourrait comprendre, mais ils estiment

qu'ils sont plus intelligents, et ils estiment davantage le raisonnement de leurs auteurs.

Ces résultats, tels quels, sont assez décevants : on aurait souhaité que les participants préfèrent des visions plus équilibrées. Ils sont également surprenants au vu du type d'argumentation utilisé dans les essais ou les articles que l'on peut trouver dans la presse : les plus longs contiennent tous des arguments allant dans la direction opposée à celle de l'auteur, et même une bonne majorité des articles les plus courts en contiennent (Wolfe, 2007). On peut cependant expliquer les résultats de Baron par un artefact expérimental dans la construction des ensembles d'arguments. Il ne s'agissait en effet que de la simple juxtaposition d'arguments prélevés parmi les réponses de participants. Ces réponses n'étaient pas du tout intégrées. Cela peut ne pas être vraiment dérangeant lorsque tous les arguments vont dans le même sens : on devine alors facilement où l'auteur fictif (puisqu'il s'agit d'arguments de différents participants additionnés) veut en venir. Mais lorsqu'il s'agit d'arguments discordants, cela peut laisser perplexe. Lorsqu'une personne présente des arguments allant dans des directions opposées, on s'attend à ce qu'il synthétise à un moment donné. Si elle ne le fait pas, elle risque d'apparaître confuse ou indécise. C'est ce qu'on dû évaluer les participants de l'expérience de Baron qui étaient confrontés à ce type d'assemblage par trop artificiel.

Dans leur expérience, Stanovich et West on fait un effort d'intégration un peu plus grand, mais il me semble que les arguments présentés n'échappent pas à la critique qui vient d'être faite. Qu'on en juge par les deux arguments suivants. Voici un argument 'unidirectionnel' :

> Well, some women get pregnant irresponsibly, and the fetus shouldn't have to be destroyed because of her mistake. Also, aborting a fetus is preventing someone from having a life, and this is wrong. None of us would have wanted to have been aborted ourselves. I worry about abortion because there is no clear place to draw the line between late abortions of fetuses that could survive on their own and the killing of unwanted infants. Also, condoning abortion is likely to reduce respect for human life in general, leading to decreased effort to preserve human life in other cases.

Bien qu'il n'y ait pas de conclusion explicite, tous les arguments vont dans le même sens, et ce discours n'a rien de choquant ni de difficile à comprendre. Et voici un argument 'bidirectionnel' :

> Families must be limited in today's world. If we are going to limit births, it is better to limit the births of unwanted children than limit the births of children who are wanted. Abortion is one means of preventing unwanted children from being born, when it is too late to limit them by other means. Also, women should be able to decide whether they want to go through something that affects them as much as pregnancy and childbirth do. On the other hand, I do believe that killing human beings is wrong and abortion is killing a human, even though the human is only a fetus. Also, abortion is never absolutely necessary as a means of birth control. There are lots of alternatives.

Il me semble clairement que cette conclusion « There are lots of alternatives » indique une absence de capacité de synthèse. Il semble que le participant (fictif) ne peut pas se décider et ne fait que lister des arguments sans vraiment réfléchir. Il ne dit même pas quelque chose comme 'il y a de bons arguments des deux côtés, il est dur de trancher', qui serait sûrement plus acceptable. Cela peut expliquer le résultat très surprenant de cette expérience : les participants attribuèrent des notes aussi basses aux paragraphes 'bidirectionnels' qu'aux paragraphes 'unidirectionnels' allant dans la direction opposée à leurs croyances personnelles.

Bien qu'elles ne s'inscrivent pas toutes explicitement dans la tradition du raisonnement motivé, les études qui ont été passées en revue dans cette section présentent des résultats qui concordent bien avec cette théorie. L'aspect motivationnel est manipulé par l'interaction entre les croyances préalables des participants et les conclusions des arguments qu'ils examinent : lorsqu'il y conflit, ils sont motivés pour les rejeter (ou mal les évaluer) ; dans le cas contraire, ils sont motivés pour les accepter (ou bien les évaluer). Dans les termes de la théorie argumentative, on peut dire qu'il s'agit là des intuitions initiales des participants, principalement formées sur la base de mécanismes de vérification de cohérence. Mais ce qui est plus intéressant est que cette motivation a par exemple pour effet de

pousser les participants à chercher des défauts dans les arguments (lorsqu'ils veulent mal les évaluer). Pour la théorie argumentative (de même que pour la théorie du raisonnement motivé), ceci survient après que l'évaluation initiale des arguments ait pris place : après que les participants aient défini la réponse intuitive qu'ils souhaitaient donner. On pourrait alors s'interroger sur l'utilité de cette phase : si on va rejeter un argument de toute façon, à quoi bon chercher des défauts ? Car, justement, trouver de tels défauts permet de justifier le rejet de la conclusion (ou la mauvaise évaluation de l'argument). L'aspect 'rétroactif' de la recherche des défauts dans les arguments est attesté par le simple fait que, selon la congruence des conclusions avec leurs croyances, les participants sont plus ou moins aptes à les trouver. Or les arguments précèdent les conclusions. S'ils étaient repérés durant la simple lecture des textes à évaluer, la position de la conclusion ne devrait pas faire de différence. Il semble bien au contraire que cette recherche soit dictée par la nécessité de trouver des justifications pour une réponse qu'ils souhaitent donner en premier lieu.

7.5 Evaluation biaisée et polarisation des attitudes

Nous venons de voir que les gens tendent à avoir des évaluations biaisées des arguments, notant plus généreusement ceux qui ont des conclusions plaisantes, étant plus critiques lorsque ces conclusions leur siéent moins. Parmi les conséquences possibles de ce phénomène, une a particulièrement attiré l'attention des chercheurs : la polarisation des attitudes.

On dit qu'une attitude se polarise lorsque, suite à un processus quelconque, elle devient plus extrême, qu'elle se renforce. Ceci peut survenir suite à de nombreux phénomènes, mais il en est un qui est particulièrement pertinent dans le cas présent : celui de la polarisation suivant l'assimilation biaisée d'arguments. Parmi les études sur le sujet, la plus citée est sans doute celle de Lord, Ross et Leper qui fut la première à démontrer une forme de polarisation des attitudes (Lord, Ross, & Lepper, 1979, mais voir Greenwald, 1969, pour un résultat antérieur très proche de celui-ci). La polarisation des attitudes s'obtient ici lorsque les participants sont confrontés à des arguments pro- et contre-attitudinaux, et que, se montrant particulièrement critique envers les arguments contre-attitudinaux et acceptant facilement les

arguments pro-attitudinaux, ils finissent avec une attitude plus extrême qu'elle ne l'était au début. Dans l'expérience de Lord et collègues, le sujet était la peine de mort, les participants appartenant à deux groupes, l'un lui étant très favorable et l'autre très défavorable. Tous les participants furent amenés à prendre connaissance du résultat de deux études, l'une concluant que la peine de mort était un bon moyen de dissuasion et l'autre non. Dans une première phase, les participants ne furent confrontés qu'à un court résumé de chaque étude, puis, dans une seconde phase, ils purent lire une version plus longue de chaque étude, accompagnée de critiques et de réponses aux critiques. Après avoir lu chaque compte rendu (quatre fois en tout donc, deux fois après chaque étude pour chaque phase), les participants devaient dire si leur attitude avait évoluée, soit simplement depuis cette dernière lecture, soit depuis le début. Ils devaient également noter la qualité de chacune des études dont ils prirent connaissance.

Pour ce qui est de la qualité des études, les participants attribuèrent des notes positives à celles soutenant leurs croyances et des notes négatives ou proches de zéro à celles s'y opposant. Etant donné que la méthodologie des études présentées était similaire dans les deux cas, et surtout que les résultats des études avec les deux méthodologies étaient contrebalancés entre les participants, il ne peut s'agir que d'un biais dans la façon dont les études sont évaluées. Mais le résultat le plus intéressant concerne l'évolution des attitudes des participants. Lors de la première phase, la façon dont les participants rapportèrent l'évolution de leurs attitudes était biaisée (comme on pourrait s'y attendre étant donné la façon dont les différentes études furent évaluées) mais pas complètement : bien que les études ayant des conclusions contre-attitudinales furent jugées plus sévèrement, les participants rapportèrent après la lecture de leur résumé une évolution de leurs attitudes dans le sens des conclusions de l'étude. Cette évolution fut plus marquée pour les études pro-attitudinales, si bien qu'après la lecture des deux résumés, durant la première phase, les participants rapportaient une évolution de leur attitude dans le sens de leur position initiale. Plus frappants encore furent les résultats de la seconde phase : quelques soient les résultats présentés, qu'ils soient pro- ou contre-attitudinaux, les participants rapportèrent une évolution de leur attitude dans le sens d'un renforcement. C'est-à-dire que des participants opposés à la peine de mort (par exemple) lisant une étude étant supposée montrer que la peine de mort était un bon moyen de dissuasion rapportèrent une évolution de leur attitude renforçant leur opposition à la peine de

morts. Au final, après la prise en compte des deux études, les participants rapportèrent donc une évolution de leurs attitudes dans le sens d'un renforcement de leurs attitudes initiales. Il s'agit de la première étude rapportant une telle polarisation des attitudes, mais elle a une limite (relevée, par exemple, par Taber & Lodge, 2006) : il ne s'agit pas d'une évolution des attitudes elles-mêmes, mais d'un rapport des participants sur l'évolution de leurs attitudes. Lord et al. ont dû se contenter de cette mesure à cause d'effets plafonds : les participants avaient déjà une attitude proche des extrémités de l'échelle au début de l'expérience, ce qui empêchait de mesurer une polarisation des attitudes en comparant attitudes initiales et attitudes après évaluation des arguments.

A la suite de ces expériences, plusieurs études ont été entreprises afin de démontrer d'une façon plus élégante la réalité du phénomène de polarisation. Une première série d'expériences de Miller et collègues (1993) retrouva l'effet d'assimilation biaisé : les participants évaluaient de façon plus clémente les arguments en faveur de leurs positions. Cependant, la polarisation directe, résultant d'une comparaison entre les attitudes précédant l'exposition des arguments et leur succédant, ne fut pas observée. Une autre expérience échoua à nouveau à trouver cet effet (Kuhn & Lao, 1996). On peut cependant se demander si ces deux dernières expériences avaient bien réussi à circonvenir à l'effet plafond mentionné dans l'étude de Lord et al.[59] Si ce n'est pas le cas, on ne doit pas s'étonner de cette absence de résultats. Enfin, la seule étude ayant réellement démontré un effet de polarisation fut celle de Pomerantz et al. (1995). Ils y parvinrent en mesurant de façon plus précise les attitudes des participants. Ils purent alors établir que ceux qui avaient une grande confiance dans leurs attitudes étaient bien victimes de ce phénomène de polarisation.

Enfin, on peut mentionner une réplication de l'étude originale, menée à bien afin d'évaluer l'efficacité de moyens d'atténuer les biais (Lord, Lepper, & Preston, 1984). Les auteurs répliquèrent leur première étude, en la comparant à deux conditions expérimentales. Dans la première ('ne soyez pas biaisés'), ils insistaient sur l'importance d'un raisonnement objectif :

[59] Kuhn et Lao ont bien essayé de construire des échelles permettant de contourner cette limite, mais leur manipulation n'était pas du tout convaincante, car elle aurait forcé les participants à prendre des positions très difficilement défendables.

> We would like you to be as objective and unbiased as possible in evaluating the studies you read. You might consider yourself to be in the same role as a judge or juror asked to weigh all of the evidence in a fair and impartial manner. (p.1233)

Dans la seconde (‘considérez le contraire’), ils insistèrent sur l’importance de considérer que les études dont ils vont prendre connaissance auraient pu avoir des résultats différents :

> Ask yourself at each step whether you would have made the same high or low evaluations had exactly the same study produced results on the other side of the issue. (p.1233)

Les résultats furent sans appel. Pour ce qui est de l’évaluation des études présentées, la réplication fonctionna, les participants favorisant la peine de mort jugeant plus sévèrement les études montrant qu’elle avait un effet dissuasif et moins sévèrement celles qui avaient des conclusions opposées, et vice versa pour ceux y étant opposés. Les instructions ‘ne soyez pas biaisés’ eurent un effet opposé à celui désiré : les évaluations étaient plus extrêmes, l’écart entre l’évaluation des études pour et contre la peine de mort se creusant. Dans le cadre de la théorie argumentative, il s’agit du résultat à attendre dans la mesure où les instructions poussent les gens à raisonner plus, mais ne parviennent pas à modifier la direction de leur raisonnement. Les instructions ‘considérez le contraire’, par contre, eurent l’effet souhaité, réduisant fortement l’écart entre l’évaluation des arguments pour et contre jusqu’à rendre la différence non significative. Cet effet se répercuta pour ce qui est de l’évolution des attitudes : alors que le même phénomène de polarisation s’observa dans la condition ‘ne soyez pas biaisés’, il disparut totalement dans la condition ‘considérez le contraire’.

Que peut-on conclure de cette section sur la polarisation des attitudes entraînée par une assimilation biaisée d’arguments ? Tout d’abord, toutes les études ayant mesuré la façon dont arguments pro- et contre-attitudinaux étaient évalués ont observé un effet des attitudes préalables, de telle façon que les arguments pro-

attitudinaux étaient jugés moins sévèrement que les arguments contre-attitudinaux. Il peut cependant s'agir d'un effet des croyances préalables ayant peu à voir avec le raisonnement : les participants pourraient simplement constater que la conclusion d'un argument est contraire à leurs croyances préalables, et de là attribuer une moins bonne évaluation à cet argument sans l'avoir réellement examiné. Une seule étude a examiné dans le détail la façon dont les participants traitaient chaque argument (celle de Pomerantz et al. 1995), mais les résultats en étaient peu clairs : alors que certaines variables prédisaient une plus grande attention portée aux messages pro-attitudinaux, d'autres prédisaient au contraire une plus grande attention portée aux messages contre-attitudinaux. Etant donné que la plupart des autres études ne prennent pas la peine de distinguer ces deux dimensions, il serait difficile de tenter d'étendre ces résultats.

Le second résultat majeur est le fait que les participants (surtout, ou uniquement, selon les études, ceux dont les attitudes initiales sont extrêmes) rapportent un changement d'attitude dans le sens d'une polarisation. Après qu'ils aient lus des arguments pro- et contre-attitudinaux, ils rapportent que leur attitude a évolué dans le sens d'un renforcement de sa direction initiale. Ce résultat a été répliqué dans toutes les études, mais son origine est douteuse. Comme l'ont noté Miller et al. (1993), il est fort possible qu'il ne s'agisse que d'un biais de réponse : les participants ne souhaitant pas dire qu'ils n'ont pas été influencés du tout par les arguments, et tentant de rester cohérent avec leur position initiale, indiqueraient ce changement dans le sens d'une polarisation. Cette explication est cohérente avec d'autres effets connus en psychologie sociale portant sur la désirabilité de certaines réponses. Elle permet également de réconcilier ce résultat avec l'apparente absence de réelle polarisation des attitudes (telle que mesurée par la comparaison des attitudes précédant et succédant la présentation des arguments).

Etudes sur la polarisation des attitudes dans le domaine politique

Deux études sur la façon dont les gens raisonnent dans le domaine politique illustrent plusieurs phénomènes qui viennent d'être examinés (voir Achen & Bartels, 2006, pour une étude des conséquences réelles de ces biais sur les comportements de

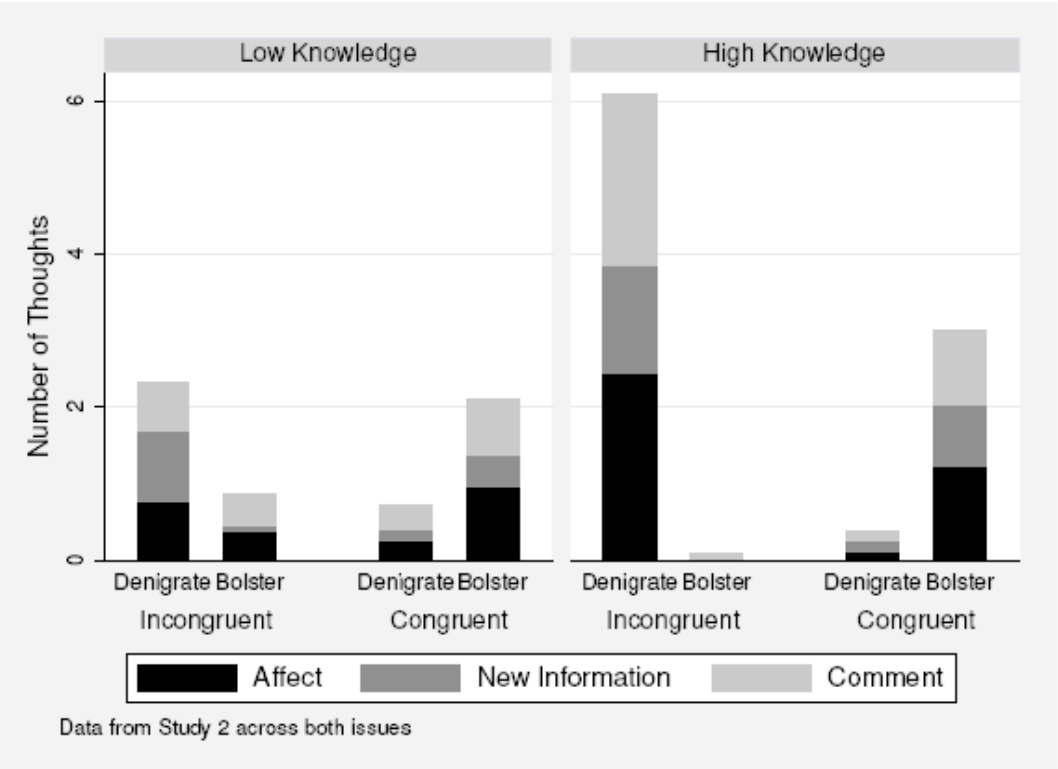
vote). Etant donné que leurs conclusions sont similaires, je me contenterai d'en détailler une, celle de Taber et Lodge, 2006, l'autre étant Redlawsk (2002).

Dans cette expérience les participants commencèrent par remplir un questionnaire sur leurs attitudes vis-à-vis de différents sujets (les deux sujets étudiés dans la suite de l'expérience étant la discrimination positive et le contrôle des armes à feu) (Taber & Lodge, 2006). Ensuite, les participants étaient confrontés à un tableau d'informations sur un de ces deux sujets. Les informations n'étaient pas visibles directement, mais les participants connaissaient leur source, ainsi que la position de cette source vis-à-vis des sujets en question. Voici un exemple d'argument utilisé : « The largest group of Americans to benefit from affirmative action thus far are women. Before 1964, women were excluded from many higher paying occupations and professions based on stereotype, custom and law. There were virtually no women police officers, lawyers, or doctors, for example. Progress has been made, but women still need affirmative action programs. » Tous les arguments étaient tirés de tracts ou de prospectus réels (puis édités). Sur les 16 éléments d'information contenus dans le tableau, les participants ne pouvaient en lire que huit. A l'issue de cette première phase, les participants remplirent à nouveau le questionnaire d'attitudes, ainsi qu'un questionnaire destiné à évaluer leur culture générale en matière de politique. Dans la seconde phase, les participants durent évaluer huit arguments (quatre pour et quatre contre) sur l'un des deux sujets utilisés ici. Leur attitude vis-à-vis du sujet en question fut de nouveau évaluée. Enfin, quatre de ces arguments (deux pour et deux contre) furent présentés à nouveau à un groupe de participant qui dut lister les pensées que ces arguments avaient évoqués lorsqu'ils avaient dû en évaluer la force.

Le premier résultat est un biais dans l'évaluation des arguments : les arguments compatibles avec les attitudes initiales des participants étaient jugés comme étant plus forts que ceux qui étaient incompatibles. Cet effet est en partie dû à un fort biais d'infirmation. Tout d'abord, les participants requirent plus de temps pour évaluer les arguments incompatibles avec leurs attitudes préalables que les autres. Qu'ils fournirent plus d'effort dans ce cas fut confirmé par le fait qu'ils énumérèrent plus de pensées pour les arguments incompatibles que compatibles. Enfin, que leur raisonnement ait été biaisé fut indiqué par le contenu des pensées : la quasi-totalité des pensées générées réfutait les arguments incompatibles et soutenait

les arguments compatibles. L'importance de cette différence est bien illustrée par la figure suivante tirée de l'article :

Figure 4: Mean Number of Thoughts for Congruent and Incongruent Arguments



Les deux tableaux représentent les participants appartenant au plus bas et au plus haut tiers pour ce qui est des réponses aux questions de culture politique. Les différentes catégories de pensées indiquées ici ne sont guères pertinentes et je ne les détaillerai pas. Mais il est frappant de voir que chez les participants ayant une certaine culture politique, pouvant être assez basique (le test étant assez simple), toutes (ou presque) les idées listées allaient dans le sens de l'attitude initiale (qu'il s'agisse de dénigrer un argument allant contre cette attitude ou de soutenir un argument allant dans son sens).

Concernant la première phase de l'expérience, dans laquelle les participants pouvaient choisir de voir des informations venant de sources s'accordant ou non avec leurs attitudes initiales, les résultats montrèrent un fort biais de confirmation : les participants choisirent une majorité d'information venant de sources avec lesquelles

ils savaient qu'ils seraient en accord[60]. Par exemple, les participants appartenant au tiers ayant la meilleure culture politique choisirent ces informations dans près de trois-quarts des cas. Ces deux phénomènes – biais de confirmation dans le choix des arguments, et biais d'infirmation dans la façon de les examiner – devraient entraîner une polarisation des croyances. C'est en effet ce qui fut observé. Les participants dans le tiers supérieur de la culture politique, ainsi que ceux ayant de fortes attitudes initiales avaient en moyenne des attitudes plus extrêmes après avoir été confronté aux arguments. On peut noter qu'il s'agit là d'une réplication des résultats de Pomerantz et al. (1995) qui avaient montré un tel phénomène de polarisation pour les participants ayant des attitudes initiales fortes.

Il faut toutefois noter une limite de ces deux études pour ce qui est des conclusions qu'on peut en tirer sur les biais du raisonnement. Dans les deux cas, on pourrait essayer de rendre compte de tous les résultats uniquement au moyen des croyances préalables des participants. Bien que ces connaissances jouent clairement un rôle prépondérant, il est très peu plausible de leur faire porter tout le blâme pour les biais dans le comportement des participants. Cela est particulièrement vrai de l'étude de Taber et Lodge. Dans cette étude, tous les biais étaient beaucoup plus importants chez les personnes ayant une bonne culture politique (ils étaient même parfois absents chez ceux ayant la culture politique la moins importante). On pourrait bien sûr imaginer que les participants puissent avoir une culture politique importante mais tellement polarisée qu'elle ne comprenne *que* des arguments allant dans le sens de leurs attitudes. Cela paraît cependant très peu probable : même en ayant des lectures partisanes, on est forcément confronté à des arguments s'opposant à nos attitudes. On sait que les articles d'opinion dans les journaux, par exemple, contiennent très souvent des arguments contre le point de vue de l'auteur – arguments qu'il réfutera généralement. Dans ce cas, il est très difficile d'expliquer uniquement par la répartition des croyances préalables un résultat tel que celui de la figure ci-dessus. On y voit que les participants ayant le moins de culture politique listaient plus de pensées réfutant des positions pro-attitudinales ou soutenant des positions contre-attitudinales que ceux ayant le plus de culture politique. Il me

[60] Il s'agit là d'un phénomène intéressant, qui rappelle ce qui a été revu dans la section sur l'exposition sélective. Ce type de biais ne peut guère s'expliquer que par le fait que les participants cherchent des informations qui pourraient être utilisées en tant qu'arguments, des informations qui renforcent, donc, leurs croyances initiales.

semble très peu probable que les participants ayant le plus de culture politique soient totalement ignorant de ces informations qui sont connues de ceux ayant moins de connaissances dans ce domaine généralement. Il est beaucoup plus naturel de penser qu'en plus d'avoir des connaissances biaisées, ces participants raisonnaient également de façon très biaisée.

7.6 Conséquences sur la persévérance des croyances

Tocqueville found it difficult to let go any notion once formed.
Hugh Brogan, Alexis de Toqueville, A Life, p.183

Le raisonnement motivé peut également contribuer à expliquer « un des phénomènes les plus robustes de la psychologie sociale » (Guenther & Alicke, 2008, p.706), la persévérance des croyances (belief perseverance). Il s'agit d'une tendance à conserver des croyances face à des preuves qu'elles sont infondées. Les premières études sur le sujet sont nées d'une inquiétude vis-à-vis de l'efficacité des débriefing pratiqués dans les expériences de psychologie sociale, particulièrement lorsqu'elles impliquent une dose de tromperie de la part de l'expérimentateur (L. Ross, Lepper, & Hubbard, 1975; Walster, Berscheid, Abrahams, & Aronson, 1967). Dans une de ces expériences, par exemple, les sujets devaient remplir une tâche et on leur donnait un feedback sur leur succès (L. Ross et al., 1975). Ensuite, on leur expliquait longuement que le feedback ne correspondait aucunement à la réalité, qu'il n'était utilisé que pour étudier les réactions des participants à des feedback positifs ou négatifs, et qu'il n'était donc en aucun cas un indicateur de la capacité des sujets à remplir la tâche initiale. Malgré ce long débriefing, l'évaluation que les participants (ou des observateurs) faisaient était fortement influencée par le (faux) feedback qu'ils avaient reçu.

L'exemple le plus frappant de persévérance des croyances vient peut-être d'une autre étude de Ross et collègues (C. A. Anderson, Lepper, & Ross, 1980). Les participants étaient confrontés à des petits textes expliquant qu'il y avait une association positive ou négative (selon les conditions) entre le fait d'aimer prendre des risques et le fait d'être un bon pompier. Les participants devaient ensuite donner leur avis sur cette relation. Cependant, certains participants avaient été débriefés : on

leur avait expliqué que ces textes avaient été fabriqués pour l'occasion, que les expérimentateurs n'avaient aucune idée de la relation réelle entre goût du risque et habilité en tant que pompier, et que maintenant les sujets devaient se prononcer entièrement sur la base de leurs croyances personnelles. Un très fort effet du texte fût néanmoins observé : les participants débriefés avaient des réponses très proches de celles des participants n'ayant pas été débriefés, et très différentes entre les conditions dans lesquelles l'association était positive et négative. Cela indique que les textes, bien qu'entièrement discrédités, ont eu une influence très forte sur les évaluations des participants. On peut expliquer cette absence de remise en cause par le fait qu'une telle remise en cause pourrait avoir un impact négatif sur l'image que les participants ont d'eux-mêmes. Après tout, dans la première partie de l'expérience, ils ont été convaincus par des arguments qui en fait n'étaient absolument pas valides. Il est donc plus confortable, dans cette situation, de continuer de penser que les arguments étaient en fait bons, ou au moins que leur conclusion était justifiée, plutôt que de s'avouer avoir été mené en bateau. Comme le notent Guenther et Alicke (2008), des phénomènes similaires ont été observés dans plusieurs domaines de psychologie sociale : recherches sur le biais de correspondance, recherches psycho-légales sur l'effet des preuves inadmissibles, recherches sur la formation des impressions entre autres (voir les références citées dans leur article, ainsi que Nyhan & Reifler, In prep. pour une illustration très parlante dans le domaine politique).

L'étude d'Anderson et collègues illustre bien le caractère fallacieux, non normatif, que peut prendre la persévérance des croyances : si une idée est totalement discréditée, nous devrions pouvoir ne plus la prendre en compte. Ceci est d'autant plus étonnant que du point de vue du fonctionnement général d'un système cognitif, on devrait s'attendre à l'inverse : généralement, les nouvelles croyances (issues de la perception) prennent le pas sur celles présentes en mémoire. Nous avons vu cependant que dans le cas des croyances communiquées, on peut s'attendre à un biais inverse (voir chapitre 4). Lorsqu'une croyance communiquée entre en conflit avec ce que nous pensions précédemment, nous devrions avoir une tendance initiale à la rejeter (tendance pouvant bien sûr être contrecarrée par la confiance accordée à la source, de bons arguments, etc.). Or les expériences portant sur la persévérance des croyances utilisent justement des informations communiquées. De plus, elles concernent généralement des domaines dans lesquels les participants n'ont pas réellement de croyances initiales, ou alors des croyances initiales faibles. Dans ce

cas, les premières informations présentées sont acceptées facilement car elles n'entrent pas en conflit avec ce que les participants pensent. Par contre, lorsque les secondes informations – discréditant les premières – sont communiquées, elles entrent en conflit avec les premières, et on doit alors s'attendre à un biais vers le rejet de ces informations (en plus du fait, déjà mentionné, que l'acceptation totale de ces secondes informations pourrait constituer un aveu d'incompétence dans l'évaluation des premières informations).

Cette explication prédit que si la seconde information présentée, bien qu'incohérente avec la première, est par ailleurs cohérente avec d'autres croyances des participants, ils seront alors capables de remettre la première en cause. Une expérience récente fournit une belle illustration de ce phénomène (Guenther & Alicke, 2008). Les participants étaient confrontés à une simple tâche d'identification de mot (lire un mot flashé très rapidement à l'écran), mais on leur avait dit que cette tâche corrélait bien avec les mesures d'intelligence générale. A la fin de la tâche, ils recevaient un feedback, positif ou négatif, mais sans lien avec leur performance réelle. Enfin, ils étaient débriefés et on leur disait que le feedback ne correspondait nullement à la réalité, qu'il ne s'agissait que d'un artifice expérimental. A chacune de ces trois étapes (avant feedback, après feedback, et après débriefing), les participants devaient s'auto-évaluer (par rapport à la tâche en question uniquement). Les résultats pour les deux premières évaluations sont logiques : aucune différence pour la première (les participants n'ayant reçu aucun feedback), suivi par une augmentation pour ceux ayant reçu le feedback positif, et une diminution lorsqu'il était négatif. Ce qui est intéressant est la troisième étape. Le phénomène de persévérance des croyances est bien observé dans le cas du feedback positif : les participants continuent d'être fortement influencés par le feedback positif, même après qu'il soit discrédité. Par contre, ils parviennent parfaitement à ne pas tenir compte du feedback négatif lorsque celui-ci est discrédité. En effet, l'information que le feedback négatif était artificiel, si elle est incohérente avec l'information précédente (le feedback lui-même), est parfaitement cohérente avec l'opinion générale que les participants ont d'eux-mêmes, et elle est donc très bien acceptée.

Le phénomène va cependant au-delà de simples phénomènes de cohérence, et on retrouve en fait la trace du raisonnement motivé. En effet, un simple mécanisme de vérification de cohérence prédirait que le feedback négatif, et donc généralement incohérent avec les croyances préalables des participants sur eux-mêmes, soit

directement rejeté. Au contraire, ce feedback eut l'effet escompté : il poussa les participants à moins bien s'auto-évaluer. Dans le cadre du raisonnement motivé, on peut expliquer ceci par le fait que les participants n'ont, à ce moment de l'expérience, pas de moyens de justifier un rejet du feedback. Par contre, lorsqu'on leur apprend que le feedback était artificiel, ils disposent d'un (bon) moyen de justifier son rejet, et ils le rejettent donc. Ce qui est intéressant, et ce qui est la marque du raisonnement motivé par rapport à un raisonnement plus objectif, est que les participants pour lesquels le feedback avait été positif ne modifièrent pas leur jugement. On peut expliquer ceci en disant qu'ils avaient à leur disposition des raisons de bien s'évaluer (le feedback positif) et des raisons de s'évaluer de façon neutre (l'invalidation de ce feedback), et que, étant motivés pour penser que leurs performances étaient bonnes, ils ne retinrent que la justification pour s'attribuer une bonne évaluation.

7.7 Effets sur la confiance, la polarisation et le renforcement

> It is astonishing what foolish things one can temporarily believe if one thinks too long alone.
>
> John Maynard Keynes, The General Theory of Employment, Interest and Money, p.vii

Nous avons vu dans l'introduction de cette section empirique que dans certains cas le raisonnement ne devrait guère avoir d'autres effets que d'augmenter notre confiance dans la justesse d'une décision, ou de renforcer une attitude. En effet, si des motivations externes nous poussent à raisonner sur une conclusion, ou une décision, dont nous n'avons pas de raisons internes de douter (nos intuitions ne sont ni faibles ni contradictoires), le raisonnement devrait trouver des justifications pour cette conclusion ou cette décision, ne faisant ainsi que la renforcer. Nous allons voir trois contextes dans lesquels de tels phénomènes ont été démontrés : la surconfiance (overconfidence), la polarisation des attitudes, et le renforcement (bolstering).

La surconfiance, le fait de surestimer les chances qu'une réponse que nous avons donnée, ou une décision que nous avons prise, soit la bonne, est un phénomène couramment observé en psychologie (voir Lichtenstein, Fischhoff, & Phillips, 1982, pour une revue des travaux initiaux). Bien que la prévalence du phénomène et ses causes soient toujours l'objet de discussions (voir par exemple Gigerenzer, Hoffrage, & Kleinbölting, 1991), ce qui nous intéresse ici est uniquement un facteur possible : le fait de penser à des raisons soutenant la décision. Les contextes expérimentaux constituent un facteur de motivation externe pour l'utilisation du raisonnement : les participants, dans la grande majorité des cas, cherchent à donner des réponses qu'ils peuvent justifier. Cependant, dans certains cas, ils auront des intuitions raisonnablement fortes sur les réponses à donner. Dans ce cas, ils devraient principalement (ou uniquement) trouver des raisons soutenant ces intuitions initiales, ce qui peut par la suite les amener à surestimer les chances qu'il s'agisse en effet de la bonne réponse.

Deux études confirment, de façon indirecte, l'existence de ce phénomène. Etant donné que la méthode globale, ainsi que les résultats, sont similaires, je les décrirai ensemble. Dans la première (Koriat, Lichtenstein, & Fischhoff, 1980), les participants devaient répondre à des questions de culture générale, alors que dans la seconde (Hoch, 1985), ils devaient estimer les chances de succès de trouver un emploi (il s'agissait d'étudiants en dernière année d'université). Dans les deux cas une manipulation identique fut effectuée : dans la condition 'pour', les participants durent générer des raisons soutenant leur réponse, dans la condition 'contre', ils devaient donner des raisons contre, et enfin dans une condition 'les deux', ils devaient fournir à la fois des raisons pour et contre. Les deux résultats les plus robustes sont les suivants : le fait de donner des raisons 'contre' tend à faire fortement diminuer la surconfiance, alors que la génération de raisons 'pour' n'a absolument aucun effet. Comme le notent Koriat et al. : « The fact that supporting instructions had no effect on performance suggests that producing a supporting reason is approximately what people normally do when asked to assess the likelihood that an answer is correct. » (p.114). Par contraste, les effets de la génération de raisons 'contre' montrent bien que les participants ne s'engagent pas spontanément dans cette activité. Ces résultats montrent donc non seulement que les participants ont bien tendance à s'engager naturellement dans une recherche biaisée de raisons

lorsqu'ils sont confrontés à des tâches expérimentales classiques, mais également que cette recherche peut les entraîner à surévaluer leur confiance dans leur réponse.

Les deux autres séries d'études que nous allons revoir se situent davantage dans le champ de la psychologie sociale. La première se penche sur les effets que peut causer le simple fait penser à un objet (au sens large : il peut s'agir d'une personne, d'un objet physique, ou d'une idée) sur l'attitude vis-à-vis de cet objet. Depuis le début des années 1970 Tesser défend l'idée que cette simple action peut créer un phénomène de polarisation : penser à un objet pourrait nous faire avoir une attitude plus extrême à son égard (voir par exemple Tesser, 1978). Tesser et ses collègues ont démontré l'existence de ce phénomène à plusieurs reprises. Les expériences et leurs résultats étant similaires, je me contenterai de décrire la première plus en détail. Dans cette expérience (Sadler & Tesser, 1973), les participants pensaient prendre part à une étude sur la façon dont on évalue les personnes en l'absence d'indices visuels. Ils étaient amenés à écouter une personne (un complice) parler pendant deux minutes. Ce complice, qui était en fait un acteur, était très plaisant dans une condition et très déplaisant dans l'autre. Ensuite, les participants devaient soit penser à la personne qu'ils venaient d'écouter, soit se consacrer à une tâche distractive, avant finalement de remplir un questionnaire sur cette personne. Les résultats attendus furent observés : non seulement l'acteur fut efficace pour créer des impressions plus ou moins positives, mais ces impressions furent plus extrêmes lorsque les participants passaient du temps à réfléchir à la personne. Utilisant une méthodologie globalement similaire, Tesser et Conlee (1975) répliquèrent ces résultats avec des sujets d'attitudes classiques ('est-ce que la prostitution devrait être légalisée ?'), en ajoutant une variable temporelle : plus les participants devaient réfléchir longtemps sur le sujet, plus leurs attitudes tendaient à se polariser.

A la suite de ces expériences Tesser, ainsi que d'autres auteurs, ont étudié les facteurs pouvant modérer cet effet de 'polarisation créée par la pensée' ('thought induced polarization'). Le premier pourrait être qualifié de « rappel de la réalité » ('reality check') : il s'agit d'une atténuation de la polarisation due à un rappel de la réalité 'objective' (Tesser, 1976). Dans cette expérience, les participants commençaient par évaluer les photographies de plusieurs tableaux. Ensuite, ils devaient en réévaluer deux après avoir été soumis à l'une des trois conditions suivantes : penser aux tableaux en leur absence, penser aux tableaux en présence de

leurs photos, et tâche de distraction. Enfin, leur opinion sur ces tableaux fut mesurée à nouveau. Dans ces circonstances, alors que de très faibles taux de polarisation furent observés suivant la tâche distractive, ils étaient plus importants lorsque les participants pensaient au tableau en le voyant, et encore plus importants lorsqu'ils y pensaient sans le voir. Dans le présent cadre, ce résultat est intéressant car il met précisément le doigt sur l'aspect métareprésentationnel des biais du raisonnement. A l'issue de la revue des travaux sur le biais de confirmation en psychologie du raisonnement, j'avais souligné le fait que certaines stratégies adéquates lorsqu'il s'agit de traiter directement du monde peuvent ne plus l'être lorsqu'il s'agit de traiter de représentations. Il semble qu'on observe ici un tel phénomène. Lorsque les participants ne peuvent compter que sur leurs pensées pour réfléchir à un objet, il y a un fort risque que celles-ci se regroupent par valence : si on est positivement biaisé vers un objet, ses aspects positifs seront plus saillants, ça sera à eux que l'on pensera en premier. De plus, si l'on pense malgré tout à des aspects négatifs, nous serons tentés de trouver des raisons pour les ignorer, car c'est là ce que fait le raisonnement. Par contre, si l'objet se trouve en face de nous, il devient plus dur d'ignorer ses aspects négatifs : d'une certaine manière, l'objet agit comme un partenaire de conversation présentant des contre-arguments.

Tesser avait cependant observé un résultat étrange dans l'expérience précédente : les résultats décrits ne s'appliquaient qu'aux femmes, ceux des hommes ne variant pas dans les différentes conditions. L'hypothèse qu'il avança est la suivante : les femmes (de cet échantillon) auraient plus de connaissances dans le domaine artistique, et auraient donc été plus à même d'élaborer et de trouver des pensées, des arguments, soutenant leur première évaluation, ce qui est la cause du phénomène de polarisation. Afin de tester cette hypothèse, il mit au point une expérience dans laquelle les participants des deux sexes devaient évaluer des stimuli relevant de la mode ou du football américain (Tesser & Leone, 1977). Les résultats confirmèrent les prédictions : alors que les attitudes des femmes ne se polarisèrent que dans le cas de la mode, celles des hommes ne le firent que pour le sport (voir aussi Millar & Tesser, 1986).

Dans la tradition de recherche inaugurée par Tesser, il est important de mentionner deux études de Chaiken et ses collègues car elles montrent bien l'importance du raisonnement dans le processus, ainsi que les effets des intuitions initiales. Dans la première (Chaiken & Yates, 1985), les participants avaient

préalablement été testés afin qu'on connaisse bien leurs attitudes vis-à-vis de la peine de mort ou de la censure. Ce test permit aux auteurs d'établir ce qu'ils nomment la cohérence affective-cognitive des attitudes : il s'agit d'un indice reflétant la concordance des différentes façons de mesurer les attitudes. Les participants retenus avaient des attitudes soient très cohérentes, soit très peu cohérentes. Lors du test en lui-même, les participants durent tout d'abord remplir un questionnaire d'attitude, comprenant les attitudes visées (afin de s'assurer qu'elles n'avaient pas évolué depuis le pré-test, et de fournir un point de comparaison direct pour mesurer la polarisation) et des distracteurs. Ensuite, ils durent écrire pendant sept minutes un essai sur un des deux sujets (peine de mort ou censure), remplir à nouveau le questionnaire d'attitude, écrire un essai sur l'autre sujet, et enfin remplir une dernière fois le questionnaire d'attitude. Les résultats portant sur la polarisation confirmèrent les conclusions des études précédentes : seuls les participants ayant écrit un essai sur un sujet à propos duquel ils avaient, dès le début, une attitude très cohérente polarisèrent leurs attitudes. Un intérêt de cette méthodologie est qu'elle permet l'analyse des essais des participants. Conformément aux prédictions, le phénomène de polarisation s'explique par le contenu des essais : les essais des participants ayant des attitudes très cohérentes comprenaient plus d'arguments défendant leur point de vue ou attaquant le point de vue adverse, et moins d'arguments allant dans la direction opposée, ainsi que d'énoncés neutres (différences significatives ou marginalement significatives). Ceci montre bien que la polarisation résulte d'un recours à des arguments biaisés, et donc du fonctionnement orienté du raisonnement (ce lien est soutenu par les études de corrélation entre le contenu des essais et la polarisation)[61]. De plus, on ne peut pas dire qu'il s'agisse ici d'une différence d'effort car le nombre d'arguments avancés par les participants n'était pas différent, quelque soit la consistance de leurs attitudes : il s'agissait donc bien d'un cas de recherche biaisée.

La seconde étude de Chaiken (Liberman & Chaiken, 1991) vise à nouveau à tester l'idée que la consistance des représentations sous-tendant les attitudes était une condition essentielle à leur polarisation. Cette fois, la consistance (ou plutôt

[61] La démonstration ne s'applique directement que pour cette expérience mais, étant donné les similarités dans les procédures, il est plus économique de l'utiliser pour rendre compte également des résultats des autres expériences.

l'inconsistance) est mesurée par les positions que prennent les participants sur des valeurs pertinentes à la réflexion sur certaines attitudes, valeurs qui peuvent se trouver en conflit. Par exemple, l'avis sur le problème de l'ouverture du courrier privé par la CIA peut être influencé par des valeurs de liberté et de sécurité. Ayant mesuré les positions des participants vis-à-vis de différentes valeurs, puis ayant procédé à une manipulation similaire à celle de l'expérience précédente (mesure des attitudes, écriture d'un essai sur l'attitude cible ou un distracteur, nouvelle mesure des attitudes), les chercheurs observèrent des résultats conformes aux prédictions : seuls les participants dont les valeurs n'étaient pas en conflit, et qui durent rédiger un essai sur l'attitude cible, virent leurs opinions se polariser.

Finalement, on peut mentionner une étude plus ancienne de Jellison et Mills (1969) qui montre qu'un des facteurs pouvant pousser naturellement au raisonnement – et à la polarisation – est l'anticipation d'une prise de position publique. Dans cette expérience, certains participants pensaient qu'ils allaient devoir enregistrer leurs opinions sur cassette, alors que d'autres n'avaient pas cette tâche à accomplir. Bien que rien d'autre ne distingue ces deux groupes, les participants du premier avaient des opinions plus extrêmes que ceux qui ne s'attendaient pas à devoir enregistrer leurs opinions. Comme le disent les auteurs : « This could have motivated them to think of additional reasons why their side was correct, to think of additional arguments favoring their side and opposing the other side. » (p.346). Les résultats qui viennent d'être passés en revue étayent cette conclusion. Il s'agit donc d'une nouvelle illustration du fonctionnement biaisé du raisonnement, bien que dans ce cas, les participants devant préparer une intervention publique, on peut être moins surpris qu'il soit justement biaisé vers la rechercher d'arguments.

D'autres études se sont penchées sur un phénomène proche de la polarisation : le renforcement (bolstering) des attitudes (W. J. McGuire, 1964). Cet effet est principalement observé lorsque des participants se sont déjà engagés sur une position donnée : dans ce cas, « la justification de cette position devient une fonction majeure de la pensée » (Tetlock, Skitka, & Boettger, 1989, p.634). Plusieurs expériences ont démontré l'existence de cet effet. Dans une manipulation par ailleurs identique à celles qui viennent d'être passés en revue, et qui provoquent une polarisation des attitudes, Millar et Tesser (1986) ont manipulé le degré d'engagement des participants. Alors que certains pensaient que la première mesure

de leur attitude serait très importante, d'autres pensaient au contraire qu'elle ne serait pas analysée du tout. La polarisation fut beaucoup plus forte dans le premier groupe, montrant que les participants s'étant engagés vis-à-vis de l'expression de leur attitude étaient plus biaisées lorsqu'ils y pensaient par la suite (voir Sears, Freedman, & O'Connor, 1964, pour un résultat similaire).

D'autres expériences ont été conduites dans l'objectif spécifique de tester ce phénomène de renforcement suite à l'engagement. Tetlock et ses collègues (1989) étudièrent les biais dans les pensées des participants vis-à-vis de divers sujets, en faisant varier deux paramètres : est-ce qu'ils fournissaient ces pensées avant ou après avoir donné leur opinion (par le biais d'un questionnaire d'attitude) sur le sujet, et est-ce qu'ils devraient rendre des comptes. Les participants ayant donné leur opinion avant eurent plus de pensées soutenant leur position que ceux qui ne devaient donner leur opinion qu'ensuite. De plus, les participants devant rendre des comptes tendaient à être plus consistants (et donc plus biaisés) que les autres, dans la mesure où les opinions du public étaient connues. Une autre étude fournit des résultats similaires : par rapport à des participants devant simplement donner une réponse privée, des participants devant énoncer leur opinion en public défendirent une attitude plus extrême (Lambert, Cronen, Chasteen, & Lickel, 1996). Il semble donc bien que lorsque les gens ont commencé de s'engager dans une direction, le raisonnement sera d'autant plus biaisé que cet engagement était public (voir également Fox & Staw, 1979 et Conlon & Wolf, 1980, pour une application au cas des coûts irrécupérables, et Lerner & Tetlock, 1999, pour une revue des effets de devoir rendre des comptes sur le renforcement). Ce phénomène semble même avoir des effets proactifs : le simple fait de savoir que l'on va s'engager publiquement tend à polariser nos pensées, en préparation des arguments qui seront avancés.

Les résultats passés en revue ici concordent avec ceux portant sur la polarisation des groupes. Les effets sont les mêmes : polarisation des attitudes, augmentation de la confiance, déplacement vers les extrêmes. Et, dans le cadre de la théorie argumentative, les causes sont les mêmes : des usages inappropriés du raisonnement. Il peut s'agir d'usages purement individuels, comme lorsqu'on demande aux participants de réfléchir à un sujet, ou en groupe, lorsque des personnes débattent d'un sujet qui, normalement, ne ferait pas débat. Il s'agit donc encore d'un cas dans lequel il est difficile d'expliquer les effets – clairement non normatifs – du

raisonnement dans le cadre des théories classiques, alors que la théorie argumentative se trouve au contraire en bonne position pour en rendre compte.

7.8 Un raisonnement objectif est-il possible ?

Pour la théorie du raisonnement motivé telle que présentée par Kunda : « the motivated reasoning phenomena under review fall into two major categories: those in which the motive is to arrive at an accurate conclusion, whatever it may be, and those in which the motive is to arrive at a particular, directional conclusion » (Kunda, 1990, p.480). On retrouve une idée similaire chez Kruglanski : « The individual's motivation to generate alternative hypotheses is assumed to be affected by needs in three separate classes: (1) the need for structure; (2) the fear of invalidity, and (3) the need for specific conclusions (the need for conclusional contents). » (Kruglanski & Freund, 1983, p.450) et « The fear of invalidity stems from the perceived costs of committing a judgmental error. Opposite to the need of structure, the fear of invalidity is assumed to have a facilitating influence upon the hypothesis-generation process because of the expected dangers of committing oneself to a given, possibly mistaken hypothesis. » (Ibid., p.450). Pour la théorie argumentative cependant, il n'est pas possible de se départir aussi aisément des biais du raisonnement, de son fonctionnement 'motivé' : il s'agit là de son seul mode de fonctionnement possible. Comment départager les deux théories ? Pour Kunda ou Kruglanski, les participants disposeraient de deux modes de raisonnement, relativement distincts, un premier qui leur permettrait de trouver des arguments soutenant les conclusions qu'ils sont motivés pour adopter, et un second qui serait capable d'investir de l'énergie pour parvenir à la meilleure réponse, à une réponse objective. Si tel était le cas, tout ce qui pousse les participants à utiliser ce second mode de raisonnement devrait augmenter les performances. Pour la théorie argumentative au contraire, il n'existe qu'une seule capacité de raisonnement, qui est fondamentalement biaisée. Si on peut s'attendre à des différences dans son fonctionnement selon les circonstances (dans le type d'arguments qui sera jugé acceptable, dans le nombre de positions qui sont envisagées, etc.), il n'en reste pas moins que rien ne garantit qu'une plus grande motivation mène à de meilleurs résultats. Au contraire, dans certains cas plus de

raisonnement devrait amener à plus de biais, même chez des participants souhaitant donner une réponse objective.

De nombreux travaux, dont certains ont déjà été examinés ici, soutiennent la thèse de la théorie argumentative. Dans les tâches de psychologie du raisonnement classique, le raisonnement motivé ne devrait guère intervenir : les énoncés sont abstraits, les croyances ou motivations des participants ne devraient pas entrer en jeu. Au contraire, les instructions mettent l'accent sur le besoin d'une réponse logique, normative. Malgré cela, on observe là aussi le même fonctionnement biaisé du raisonnement, qui va principalement chercher à soutenir les intuitions initiales des participants et non à donner une réponse objective. Quels sont les effets de la motivation dans ces circonstances ? Ils sont limités. Il arrive que des instructions insistant particulièrement sur la nécessité de donner une réponse logique améliorent quelque peu les performances – sans en changer les patterns fondamentaux (Evans, 2002). Dans d'autres cas, ces mêmes instructions échouent à amener un réel changement de stratégie chez les participants. Ainsi, dans le cas du test d'hypothèse, les participants peuvent adopter une stratégie de test négatif, mais pas une stratégie de falsification – ce qui leur est pourtant demandé (Poletiek, 1996; Tweney et al., 1980). Enfin, dans une tâche au moins (la tâche de sélection), des promesses de gain, qui pourtant devraient augmenter la motivation de donner la bonne réponse, n'ont aucun effet (Johnson-Laird & Byrne, 2002; Jones & Sugden, 2001).

Si on se tourne maintenant vers les tâches de jugement et de prise de décision, il semble que les effets de la motivation soient encore moins importants. J'ai déjà mentionné plusieurs revues qui indiquent que les promesses de gain n'ont, dans la grande majorité des cas, pas d'effet sur ce genre de tâche – un résultat vraiment surprenant si les gens étaient capables d'utiliser un 'raisonnement objectif' sur commande (Ariely et al., In Press; Arkes, Dawes, & Christensen, 1986; S. E. Bonner et al., 2000; S. E. Bonner & Sprinkle, 2002; Camerer & Hogarth, 1999). Pour ce qui est de l'effet des instructions, les résultats d'une série d'expérience de Pelham et Neter (1995) sont particulièrement informatifs. Les auteurs ont comparé les effets de différentes instructions sur des problèmes classiques de jugement impliquant l'utilisation de la loi des grands nombres (le problème des hôpitaux de Kahneman & Tversky, 1972, et un autre problème similaire). Alors qu'un jeu d'instructions disait aux participants que l'objectif de l'étude était d'évaluer les relations entre jugement et intuition, un autre mettait l'accent sur les relations entre les résultats à ce test et les

capacités d'intelligence, disant que les chercheurs étaient intéressés par les performances des participants. De plus, une version simplifiée des problèmes fut créée pour l'occasion. Dans sa version initiale, moins facile, les participants qui devaient être les plus motivés pour donner la bonne réponse (ceux ayant reçu le second jeu d'instructions) eurent de moins bons résultats. Par contre, ils furent meilleurs pour la version facile du problème. Un résultat similaire fut observé pour deux autres tâches de jugement : les participants les plus motivés avaient de moins bonnes performances pour les tâches plus difficiles, et de meilleures performances pour les tâches plus faciles. Igou et Bless (2007) démontrèrent un effet similaire pour les effets de cadrage : les participants les plus motivés (ceux qui passaient le plus de temps sur la tâche, ou qui devaient rendre des comptes) étaient plus victimes des effets de cadrage que les autres.

Si on s'intéresse maintenant aux tâches dans lesquelles les participants sont naturellement motivés pour donner une certaine réponse, on s'aperçoit que les instructions mettant l'accent sur l'objectivité peuvent avoir un effet contraire et augmenter les biais des participants. C'est ce à quoi on peut s'attendre si les instructions ont pour effet de faire raisonner les participants davantage, mais que ces derniers ne peuvent se détacher d'une utilisation biaisée du raisonnement au profit d'une utilisation plus objective. J'ai déjà mentionné les résultats d'une tentative de réduction des biais de Lord et al. (1984) qui avait eu les effets opposés : les participants instruits de l'importance d'évaluer des arguments en toute objectivité furent encore plus biaisés dans leurs jugements que ceux d'une condition contrôle. Un résultat similaire (bien que non significatif) fut obtenu par Frantz et Janoff-Bulman (2000) : des participants motivés pour se montrer justes dans leur évaluation furent marginalement plus biaisé dans leur jugement d'un conflit entre deux personnes. Au vu de ces résultats, on peut se demander si, ironiquement, il n'est pas possible que les instructions que l'on rencontre fréquemment mettant l'accent sur l'objectivité ne soient pas en partie responsables des biais qui sont observés. Ainsi, quand Taber et Lodge (2006) notent que l'effet des croyances initiales est « systématique et robuste parmi les participants sophistiqués et ceux qui ont les opinions les plus fortes, malgré nos importants efforts pour motiver l'objectivité », il est possible qu'en fait l'effet des croyances initiales soit aussi systématique et robuste non pas malgré, mais bien *à cause* des efforts faits pour motiver l'impartialité. Finalement, on peut mentionner deux d'expériences de Molden et

Higgins portant sur les variations interpersonnelles dans la motivation pour différents types de raisonnement (Molden & Higgins, 2008). Les stratégies qu'ils nomment 'vigilantes' sont celles qui se rapprochent le plus du raisonnement objectif décrit par Kunda, et pourtant… ce sont les participants qui les préfèrent qui furent les plus biaisés, montrant une préférence plus forte pour des explications internes de leurs succès et externes de leurs échecs.

A nouveau, l'argument n'est pas qu'être motivé pour donner une bonne réponse mène nécessairement à une baisse de performance. La théorie argumentative peut parfaitement s'accommoder des résultats présentés par Kunda, ou d'autres, montrant que dans certaines circonstances les gens sont, en effet, capables de donner de meilleures réponses lorsqu'ils sont motivés pour le faire. Ce qui est important c'est que la volonté des participants ne suffit pas. Ils ne peuvent pas décider de passer dans un 'mode de raisonnement objectif' qui leur garantirait une meilleure réponse, à tout le moins une réponse objective. Au contraire dans certains cas, et avec la meilleure volonté du monde, il arrivera que raisonner davantage ne change rien, ou ait même des effets négatifs. Si les participants peuvent choisir de raisonner plus, il leur est beaucoup plus difficile de choisir de raisonner mieux, si une telle chose est même possible. Les circonstances peuvent modifier la façon dont fonctionne le raisonnement, mais ceci ne dépend généralement pas de la volonté des participants. S'il est indéniable que le raisonnement peut parvenir à des résultats objectifs, qu'il puisse le faire par des moyens non biaisés reste à démontrer.

8 Raisonnement et prise de décision

Il n'est pas facile de faire des prédictions générales concernant les performances du raisonnement au-delà de sa fonction propre, la recherche et l'évaluation d'arguments. Nous avons déjà vu qu'il remplissait bien cette tâche, nous permettant d'évaluer des arguments lorsque nous sommes motivés pour le faire, ainsi que de former des arguments convaincants. Il permet par exemple aux groupes confrontés à des tâches de raisonnement de converger vers la bonne réponse par la soumission et l'examen d'arguments. Le raisonnement, cependant, est aussi souvent utilisé lors de la prise de décision. S'il est toujours employé pour trouver des arguments, ces arguments sont alors utilisés pour convaincre qu'une décision est bonne. Deux questions différentes se posent alors : d'une part, est-ce que le raisonnement continue de bien accomplir sa fonction ? Est-ce que les arguments qu'il trouve en faveur de l'option qui est finalement choisie sont bons ? Il s'agit malheureusement d'un aspect sous-étudié du processus de prise de décision, mais nous verrons que les quelques éléments disponibles pointent vers une assez bonne efficacité du raisonnement dans ce domaine.

La seconde question est : est-ce que le raisonnement mène à de meilleures décisions d'un point de vue plus général ? Le problème est alors d'établir les critères qui seront pertinents pour juger de ce qui est une bonne décision. Du point de vue le plus global, le fait de pouvoir être facilement justifiée donne un certain avantage à une décision (au moins socialement). Cet élément devrait donc faire partie des critères permettant de décider si une décision est bonne. Mais si on inclut ce critère, en ajoutant la prémisse que l'esprit est globalement bien construit, on arrive à la conclusion que les décisions prises avec l'aide du raisonnement tendront en effet à être meilleures. Elles devraient l'être ne serait-ce que parce qu'elles sont plus aisément justifiables. Mais dans ce cas, les décisions s'approchent toujours de l'optimalité, et on ne peut donc pas vraiment juger de leur adéquation avec d'autres critères. On s'intéresse plutôt ici à l'effet qu'a cette recherche de justification sur les autres aspects de la décision : est-ce que la décision est meilleure 'en elle-même' ? C'est justement cela qui est examiné dans la plupart des travaux portant sur la prise de décision, mais le problème des critères se pose encore. Certains adoptent des critères normatifs tels que ceux de la théorie des probabilités ou des théories de la

décision. Des critères normatifs moins stricts (est-ce qu'un choix est vu comme étant le meilleur par des experts ?) peuvent également être utilisés. D'autres enfin ont recours à des critères de satisfaction personnelle : est-ce que le choix fait à un moment t s'avère satisfaisant à un moment $t+1$?

De nombreuses expériences ont tenté de répondre à ces questions. Ces travaux sont intéressants pour nous car pour les théories classiques, le raisonnement devrait permettre de prendre de meilleures décisions, au moins lorsqu'il s'agit de situations relativement nouvelles. On va voir que ce critère vague pourrait être rempli par toutes, ou presque, les expériences qui vont être présentées. A l'inverse, la théorie argumentative ne prédit pas du tout une telle amélioration uniforme.

8.1 Expliquer nos choix

Les travaux de Wilson et collègues sur le lien entre attitude et comportement

A partir des années 80, Timothy Wilson et ses collègues se sont intéressés aux effets de l'introspection. L'introspection consiste ici à demander aux participants de penser aux raisons de leurs choix ou aux raisons pour lesquelles ils ont telle ou telle attitude. Dans les premiers travaux, la mesure de performance utilisée est celle du lien entre attitude et comportement. Les participants doivent donner leur attitude vis-à-vis de différents objets, et leur comportement vis-à-vis de ces mêmes objets est ensuite mesuré. Dans certains cas, les participants devaient donner des raisons pour leurs attitudes. De façon générale, cela eut pour effet de diminuer la corrélation entre attitude et comportement (voir Wilson, Dunn, Kraft, & Lisle, 1989).

Il peut sembler étrange d'utiliser une telle mesure comme un indice de performance du raisonnement. On peut donner deux arguments. D'une part, et au moins pour la théorie présentée ici, c'est bien de raisonnement qu'il s'agit lorsqu'on demande aux participants de donner des raisons pour un choix, une attitude, etc. D'autre part, un échec d'introspection peut bien être considéré comme un échec. Lorsqu'on demande aux participants de donner leur attitude vis-à-vis d'un objet, s'ils donnent une réponse qui n'est pas du tout en phase avec leur comportement suivant, cela signifie qu'ils se sont avérés incapables de comprendre, d'inférer, ou d'accéder à

leur propre attitude. Or, il s'agit là de la tâche à accomplir. Il s'agit d'autant plus d'une forme d'échec que, comme nous allons le voir, la capacité d'évaluer des attitudes qui, en effet, prédisent le comportement est supérieure lorsque les participants n'utilisent *pas* le raisonnement.

Dans la première des expériences visant à explorer spécifiquement ce phénomène, Wilson et ses collègues ont observé le comportement de participant face à des puzzles (Wilson, Dunn, Bybee, Hyman, & Rotondo, 1984). Les participants commençaient par résoudre différents puzzles, puis ils devaient les évaluer. Dans une condition (la condition 'raisons'), on demandait aux participants (avant qu'ils ne commencent) de se concentrer sur les raisons pour lesquelles ils avaient telle ou telle opinion des différents puzzles. Après avoir résolu différents types de puzzles, on donnait une feuille aux participants de cette condition afin qu'ils notent ces raisons. Ceux de la condition contrôle devaient, eux, remplir une tâche distractive (des questions démographiques) de même durée. Les participants devaient ensuite évaluer les différents puzzles. Finalement, l'expérimentateur prétendant devoir s'occuper d'un autre participant un court instant, ils étaient laissés seuls dans une pièce avec des puzzles nouveaux mais appartenant aux types déjà rencontrés, alors qu'un autre expérimentateur les observait subrepticement à travers un miroir sans teint.

Les résultats furent clairs. Dans la condition contrôle il y avait une corrélation significative entre les notes attribuées aux puzzles et le temps passé à jouer avec eux lorsque les participants étaient laissés seuls. Par contre, cette corrélation était significativement inférieure, et non significativement différente de zéro, dans la condition 'raisons'. De plus, les données indiquent que cette différence est due entièrement au fait que les participants de la condition 'raison' rapportèrent des attitudes différentes. En effet, il n'y eut aucune différence dans le comportement des participants entre les conditions, alors que l'intérêt rapporté pour les différents puzzles était lui significativement différent. Enfin, plus les participants donnèrent de raisons pour justifier leurs évaluations, moins ces évaluations corrélaient avec le comportement.

Ces résultats s'avérèrent très robustes et furent répliqués (Millar & Tesser, 1989) et étendus à plusieurs situations très différentes : jugements esthétiques (Wilson et al., 1984, expériences 2 et 3), jugements vis-à-vis de boissons ou de personnes (Wilson, Kraft, & Dunn, 1989), et même évaluation de leur vie de couple

par les participants (Wilson et al., 1984, expérience 4) (voir Wilson, Dunn et al., 1989, pour une revue plus complète, et Levine, Halberstadt, & Goldstone, 1996, pour une étude plus précise de l'effet de donner des raisons).

Extension à d'autres domaines

On pourrait arguer que les résultats qui viennent d'être passés en revue ne représentent qu'un aspect très limité des performances du raisonnement. Heureusement, Wilson et ses collègues ont appliqué des méthodes similaires à d'autres domaines.

Dans une première expérience les performances des participants étaient jugées à l'aune des évaluations de différents produits faites par des experts (Wilson & Schooler, 1991). Les participants devaient goûter cinq confitures différentes. Dans la condition 'raison', on leur demandait pourquoi ils pensaient ou ressentaient telle ou telle chose vis-à-vis de chaque confiture. Entre le moment où ils goûtaient les confitures et celui où ils les évaluaient, ils devaient remplir une feuille avec ces raisons (afin, leur était-il dit, d'organiser leurs pensées, cette feuille n'étant pas ramassée par l'expérimentateur). Dans la condition contrôle, ils remplissaient à la place une feuille dans laquelle ils devaient donner des raisons pour un choix sans rapport avec l'expérience. Les évaluations des participants furent ensuite comparées avec les évaluations faites par des experts. Alors que la corrélation était positive et significative dans la condition contrôle, elle était significativement moindre et non significativement différente de zéro dans la condition 'raison'.

Une seconde expérience utilisa ensuite une mesure différente de la performance : la satisfaction des participants eux-mêmes suivant le choix (Wilson et al., 1993). La méthodologie était similaire à celle de l'expérience qui vient d'être décrite, à la différence que les objets évalués étaient cette fois des posters. De plus, à l'issue de l'expérience, les participants pouvaient choisir (de façon totalement privée) un des posters et le conserver. Enfin, à la fin du semestre, les participants furent appelés et leur satisfaction vis-à-vis du choix qu'ils avaient fait fut évaluée. Bien que, lors de l'expérience, les participants des conditions contrôle et 'raison' évaluèrent

aussi favorablement le poster qu'ils choisirent, lorsqu'ils furent rappelés plusieurs semaines plus tard, les premiers s'avérèrent être significativement plus satisfaits de leur choix que les seconds. Ici encore, le fait d'analyser les raisons pour lesquelles ils ressentaient tel ou tel sentiment vis-à-vis des posters avait conduit les participants à faire de plus mauvais choix.

Enfin, une dernière expérience dans cette série par Wilson et ses collègues permit d'examiner la capacité des participants à prédire leur propre comportement (Wilson & LaFleur, 1995). Les participantes étaient des jeunes femmes appartenant à une sororité. Elles devaient tout d'abord choisir deux autres membres de leur sororité qu'elles avaient récemment rencontrés : celui dont elles avaient eu l'impression la plus positive, et celui dont elles avaient eu l'impression la plus négative. Ensuite, elles durent évaluer la probabilité qu'elles se comportent de certaines façons vis-à-vis de ces personnes. A nouveau, dans la condition 'raison' elles durent noter des raisons pour chacune de ces prédictions. A la fin du semestre, les expérimentateurs retournèrent dans les sororités et interrogèrent à nouveau les participantes, cette fois pour savoir si elles s'étaient comportées de la façon mentionnée dans la première partie de l'étude. Il était donc possible de comparer les prédictions aux comportements effectifs des participants. Les participantes de la condition 'raison' furent dans l'ensemble moins performantes pour juger de leur propre comportement futur que celles de la condition contrôle (71% de prédictions correctes contre 79%).

Une expérience similaire fut conduite par Halberstadt et Levine (1999). Elle concernait cette fois la prédiction de résultats de matchs de basket-ball. Les participants étaient donc pour cette expérience des étudiants connaisseurs de basket-ball. Ils devaient prédire les résultats, ainsi que la marge de victoire, de huit matchs qui allaient se dérouler dans la semaine suivant l'expérience. Dans une condition (raison), les participants devaient énoncer le plus possible de raisons expliquant leurs prédictions. Dans l'autre (intuition), il leur était au contraire demandé d'utiliser au maximum leur intuition. Les prédictions furent ensuite comparées aux résultats des matchs. Les participants de la condition 'intuition' furent significativement plus nombreux à pronostiquer la victoire du vainqueur effectif que ceux de la condition 'raison' (70% contre 65%). De même, la marge de victoire annoncée par les premiers

était plus proche de la réalité – légèrement mais significativement (11,10 contre 10,20 points).

Une vision plus équilibrée est offerte par McMackin et Slovic (2000), qui mettent l'accent sur le fait que l'analyse des raisons pourra jouer un rôle positif ou négatif selon le type de tâche. Ils eurent donc recours à une tâche 'analytique' (dans laquelle les raisons sont supposées améliorer les performances) et une tâche 'intuitive' (dans laquelle elles sont au contraire censées les détériorer). La tâche intuitive consistait à deviner l'évaluation qui avait été faite par d'autres participants d'affiches publicitaires. Des questions de culture générale (la longueur de l'Amazone par exemple) servaient de tâche analytique. Dans la condition 'raison' (par opposition à une condition contrôle), les participants étaient instruits qu'ils devaient raisonner de façon analytique, mettre de côté leurs émotions, et lister les facteurs les plus importants pour leur décision. Les résultats se conformèrent aux prédictions. Pour la tâche intuitive, les performances étaient moins bonnes dans la condition 'raison' que dans la condition contrôle. A l'inverse, pour la tâche analytique, les performances étaient supérieures dans la condition 'raison'.

On peut cependant s'interroger sur le rôle exact qu'ont joué les raisons dans la tâche analytique. Parmi les réponses aux cinq questions utilisées, quatre étaient significativement différentes entre les conditions, et toutes dans le sens d'une augmentation qui continuait de sous évaluer largement la réponse correcte. Par exemple, interrogés sur la surface des Etats-Unis, la moyenne (géométrique) des réponses était 750 399 km^2 pour la condition contrôle, 1 642 506 km^2 pour la condition 'raison' alors que la bonne réponse est 3 540 940 km^2. Dans ce cas, peut-être que les participants pensaient surtout à des raisons de donner un grand chiffre. Penser davantage à ces raisons a alors pu augmenter l'estimation donnée, mais pas pour des raisons 'rationnelles', simplement parce que plus de poids était accordé aux raisons préexistantes. Il se trouve dans ce cas que les participants tendaient à sous-évaluer la réponse correcte. Mais si leur évaluation première avait été bien calibrée, alors le fait de devoir donner des raisons aurait pu au contraire avoir des conséquences négatives. Bien que spéculative, cette analyse est soutenue par le fait que dans d'autres expériences (Koriat et al., 1980), le fait de donner des raisons n'avait pas d'influence sur la correction des réponses à des questions de culture générale. La différence qui pourrait expliquer cette disparité est qu'il s'agissait alors

de questions à choix multiples, et les raisons ne pouvaient alors pas jouer de façon quantitative comme elles le firent dans l'expérience de McMackin et Slovic. Il me semble donc hasardeux, au vu de ces résultats contradictoires, de tirer des conclusions fortes sur la base des résultats à quatre questions de culture générales qui peuvent très bien ne représenter qu'un échantillon biaisé de ce type de problème.

Depuis, le flambeau allumé par Wilson et ses collègues est passé à Ap Dijksterhuis. L'objectif de ce dernier est cependant un peu différent de celui de Wilson. Il ne s'agit pour lui pas tant de chercher à comprendre les effets que peut avoir l'analyse des raisons, mais de montrer que la 'pensée inconsciente' est souvent plus efficace que la 'pensée consciente'.

Les expériences de Dijksterhuis

Dijksterhuis défend la 'théorie de la pensée inconsciente' selon laquelle cette pensée inconsciente serait capable de traiter beaucoup plus d'informations que la pensée consciente et, partant, de trouver de meilleures solutions à des problèmes complexes. La pensée consciente, par opposition, aurait tendance à se focaliser sur quelques informations et serait donc supérieure pour résoudre des problèmes plus simples. Cette théorie fut testée par une première série d'expériences publiée en 2004 (Dijksterhuis, 2004).

Dans la première expérience, les participants devaient évaluer des appartements. Douze caractéristiques de quatre appartements apparaissaient une par une sur un écran. Chaque caractéristique était positive ou négative de façon non ambiguë. Parmi les quatre appartements, un était supérieur aux autres (huit attributs positifs et quatre négatifs), deux étaient neutres (six attributs positifs et six négatifs), et un était inférieur (quatre attributs négatifs et huit positifs). Après avoir pris connaissance de tous les attributs (48 au total), les participants étaient répartis dans une de trois conditions. Dans la condition 'réponse immédiate', les participants devaient évaluer, sur une échelle de 1 à 10, les quatre appartements. Dans la condition 'pensée consciente', les participants devaient bien réfléchir à leur avis sur les appartements pendant trois minutes avant de les évaluer. Enfin, dans la condition

‘pensée inconsciente’, cette même durée de trois minutes précédant l’évaluation était occupée par une tâche distractive. Les résultats, s’ils vont dans la direction prédite, sont assez faibles. Les participants de la condition ‘pensée inconsciente’ furent les seuls à évaluer de façon significativement différente les appartements supérieur et inférieur. Cette différence est cependant modérée par un assez surprenant effet du sexe : alors que la condition ‘pensée inconsciente’ était largement supérieure à la condition ‘réponse immédiate’ pour les hommes, il n’y eut aucune différence pour les femmes. De plus, la différence entre ‘pensée consciente’ et ‘pensée inconsciente’ ne fit qu’approcher le seuil de significativité.

La seconde expérience était très proche de la première, seul le format de présentation, contenant par exemple un tableau récapitulant les caractéristiques des appartements, était modifié. Les résultats furent globalement similaires, démontrant un effet positif de la pensée inconsciente, mais pas par rapport à la pensée consciente – uniquement par rapport à l’absence de pensée. Enfin, une troisième expérience impliquait l’évaluation de colocataires, toujours présentés par une liste d’attributs. A nouveau, des résultats proche de la première expérience furent obtenus, avec cette fois encore des résultats très différents selon le sexe des participants.

Il est difficile de tirer des conclusions fortes de ces expériences. Tout d’abord, la différence entre pensée consciente et pensée inconsciente n’est significative que dans une seule des trois expériences. De plus, dans ce cas (expérience 3) aussi bien que dans le cas ou la différence fut presque significative (expérience 1), elle est modérée par un effet du sexe tel que l’effet est totalement absent chez les femmes. On voit mal comment réconcilier ce résultat avec des conclusions très générales sur la valeur relative de la pensée consciente et inconsciente, il semble au contraire qu’il ne s’agisse que d’effets isolés, peut-être propres à un matériel ou des participants particuliers (ce que confirmeront les tentatives de réplications qui seront examinées plus bas).

Une nouvelle série d’expériences fut publiée en 2006 (Dijksterhuis, Bos, Nordgren, & van Baaren, 2006). Les deux premières expériences sont très similaires aux deux expériences rapportées ci-dessus, mais en ajoutant un facteur de complexité. Il s’agissait cette fois de voitures, toujours caractérisées par des attributs positifs ou négatifs. Cette fois, les différences étaient plus marquées, puisque l’option

supérieure possédait 75% d'attributs positifs, les deux options neutres 50% et l'option inférieure 25%. La nouvelle manipulation concerne le nombre d'attributs : dans une condition (complexe) il y en avait 12 et dans l'autre (simple) uniquement 4. Cette manipulation visait à tester l'hypothèse selon laquelle la pensée consciente serait particulièrement inappropriée pour les choix complexes. Dans la première expérience, les participants devaient choisir une voiture, alors que dans la seconde ils devaient toutes les évaluer. Enfin, dans ces expériences, seules étaient comparées des conditions 'pensée consciente' et 'pensée inconsciente'. Conformément aux prédictions, les participants de la condition 'pensée consciente' eurent des performances significativement inférieures lorsque le choix était complexe, ceux de la condition 'pensée inconsciente' ayant des résultats aussi bons dans les deux cas, qu'il s'agisse de choisir la meilleure voiture ou de lui attribuer une note supérieure. Les auteurs ne précisent cependant pas si les performances de la condition 'pensée inconsciente' était significativement supérieures aux performances de la condition 'pensée consciente' dans la condition 'complexe'.

Les deux autres résultats exposés dans cet article ne sont pas expérimentaux. Il s'agit d'enquêtes sur la satisfaction de personnes ayant acheté différents produits. Les deux variables mesurées ici sont la complexité du produit[62] et le temps que les participants ont rapporté avoir pensé au produit entre le moment ou ils l'ont vu pour la première fois et le moment ou ils l'ont acheté. A nouveau, la prédiction est que pour les objets complexes un temps de pensée plus long devrait mener à moins de satisfaction, et vice versa pour les objets simples. Dans l'étude trois, la prédiction fut vérifiée chez des étudiants qui avaient à choisir un objet qu'ils avaient récemment acheté et dire le temps qu'ils avaient passé à penser à l'achat avant de l'effectuer. Alors qu'il y avait une corrélation positive entre le temps passé à penser au choix et la satisfaction pour les objets simples, cette corrélation était négative pour les objets complexes (les deux corrélations étant significatives). Dans l'étude quatre, la complexité fut simplement contrôlée en interrogeant des gens sortant d'un magasin d'ameublement (choix complexes) ou d'une épicerie (choix simples), et à nouveau la prédiction fut vérifiée. A la sortie du magasin d'ameublement, les personnes ayant réfléchi plus que le médian étaient moins satisfaites, alors qu'elles l'étaient plus si elles avaient acheté des produits à l'épicerie.

[62] Le nombre de traits sur lequel on peut se baser pour évaluer un produit, établi dans une étude pilote.

On peut cependant critiquer ces dernières études car la complexité est liée à l'importance et au prix des produits. Lorsqu'on pense à un produite qu'on compte acheter, on a tendance à le comparer à d'autres options. Et plus on pense aux autres options, plus on aura d'opportunités de regretter le choix qu'on a fait. De nombreuses études montrent que les gens qui réfléchissent trop avant de faire un choix, qui tentent de maximiser à tout prix, sont moins satisfaits de leur choix que ceux qui ont recours au satisficing (Schwartz, 2004). Il est logique de penser que de tels regrets soient plus forts lorsqu'il s'agit d'un achat important (appareil photo) que d'un achat parfaitement trivial (shampoing). Dans le cas des achats importants, on peut très bien imaginer que les participants ayant réfléchi d'avantage à leur choix aient fait des choix qui soient aussi bons. Mais ils en viennent à être moins satisfaits uniquement car ils en savent plus sur les éventuels défauts de leur choix, ou sur les avantages qu'auraient eus d'autres choix. Dans l'ensemble ces expériences n'apportent donc qu'un soutien très limité à la théorie de la pensée inconsciente telle qu'énoncée par Dijksterhuis : les deux premières expériences ne sont pas vraiment concluantes sur la supériorité de la pensée inconsciente (pas de différence significative rapportée), et les résultats des deux études suivantes peuvent être expliqués sans faire appel à une réelle supériorité de la pensée inconsciente (voir également Dijksterhuis & van Olden, 2006).

Tentatives de réplication et méta-analyse

Sur la base de ces résultats, Dijksterhuis tire des conclusions très fortes sur la façon optimale de prendre des décisions. Ainsi, dans l'article de 2004, il donne les conseils suivants :

> When faced with complex decisions such as where to work or where to live, do not think too much consciously. Instead after a little conscious information acquisition, avoid thinking about it consciously. Take your time and let the unconscious deal with it. (Dijksterhuis, 2004, p.596)

Récemment, plusieurs groupes ont pointé la relative faiblesse des résultats sur lesquels Dijksterhuis se base pour tirer des conclusions aussi fortes. D'une part,

comme la revue de ses expériences l'a montré, dans plusieurs cas les différences entre condition 'pensée inconsciente' et 'pensée consciente' ne sont pas significatives. Dans d'autres cas elles le sont, mais sont modérées par des facteurs qui ne devraient pas être pertinents (le sexe des participants). Des tentatives de réplication ont donc été menées.

Une série de réplication fut conduite par Newell et ses collègues (In Press). Ils indiquent une limite importante des résultats de Dijksterhuis (2004) dans son étude sur le choix d'appartements. Dans les deux expériences originales, le critère retenu était l'appartement ayant le plus d'attributs positifs. Il est cependant probable que différents attributs puissent être pondérés différemment. Dans ce cas, la simple quantité d'attributs positifs peut ne pas révéler la valeur réelle des appartements, qui correspond plutôt à la valeur pondérée des différents attributs. Lorsque la pondération des attributs est prise en compte, Newell et al. ont observé que c'est elle, et non la simple quantité d'attributs positifs ou négatifs, qui jouait le rôle le plus important. De plus, dans une première expérience ils n'ont observé aucune différence entre des conditions 'pensée consciente', 'pensée inconsciente' ou 'réponse immédiate'. Dans une deuxième expérience, pour laquelle un tableau récapitulatif était utilisé afin d'imiter l'expérience 2 de Dijksterhuis (2004), c'est la condition 'pensée consciente' qui donna les meilleurs résultats. Enfin, dans une troisième expérience, l'expérience 1 de Dijksterhuis et al. (2006) fut répliquée. Cette fois encore, les résultats favorisèrent (non significativement) la condition 'pensée consciente', et ce même dans le cas des objets complexes.

Une autre réplication a récemment été conduite par Acker (2008). Cette fois, c'est l'expérience 2 de Dijksterhuis et al. (2006), qui était répliquée. Cette fois encore, la réplication échoua, et donna même des résultats inverses. Acker c'est également livrée à une méta-analyse des résultats de 17 expériences (la plupart n'étant pas encore publiées, il n'est pas possible d'en relater le détail ici) comparant pensée consciente et inconsciente. Sur les 17 études, seules cinq montraient un avantage significatif pour la pensée inconsciente, aucune ne montrant un avantage significatif dans l'autre direction. Au total, l'effet cumulé des 17 études peut-être jugé comme étant non significatif. De même, les différences entre pensée

inconsciente et réponse immédiate, et entre pensée consciente et réponse immédiate n'étaient pas significatives.

Que peut-on conclure de cet ensemble d'études de Wilson, Dijksterhuis, et leurs collègues ? Il n'est pas facile d'en tirer de conclusions positives. Les différences entre conditions sont souvent faibles, parfois à peine significatives ou complètement dépendantes de facteurs dont on ne sait pas pourquoi ils jouent un rôle. De plus, certains résultats s'avèrent impossibles à répliquer. La seule exception, le seul résultat qui semble vraiment robuste, est celui de l'examen des raisons sur la consistance entre attitude et comportement. Mais il s'agit d'un phénomène bien particulier à partir duquel il serait hasardeux de généraliser. Le résultat le plus intéressant est donc peut-être un résultat négatif : dans presque tous les cas les quelques minutes accordées aux participants pour réfléchir, dans la condition 'raisons', ou 'pensée consciente' n'entraînèrent aucune amélioration des résultats par rapport à une réponse immédiate (voir également Klein, 1998, pour des résultats similaires dans des situations de prise de décision réelles). On voit mal pourquoi le raisonnement, s'il avait bien la fonction qui lui est couramment attribuée, ne pourrait pas aider dans ce type de situation : il s'agit bien de prise de décision, dans des situations relativement nouvelles.

8.2 Le choix basé sur des raisons

You gotta have a reason, everything has a reason.
Dr Gregory House, House M.D.

A la fin des années 80 a commencé d'émerger une perspective sur la prise de décision dont les conclusions sont très proches de celles de la théorie argumentative : la théorie du '*choix basé sur des raisons*' ('reason based choice', ou 'choice based on reasons') (Shafir, Simonson, & Tversky, 1993). Selon cette théorie, les gens prennent certaines décisions car ils peuvent facilement trouver des raisons pour les défendre – et non car il s'agit de décisions maximisant leur utilité ou se conformant à telle ou telle norme des théories de la décision. Pour la théorie argumentative, c'est précisément ce qui devrait se passer lorsque le raisonnement joue un rôle causal dans

la prise de décision : il devrait avoir pour effet de nous faire prendre des décisions justifiables – qu'elles soient meilleures selon d'autres critères ou non. La théorie du choix basé sur des raisons a été appliquée à de nombreuses situations de prise de décision, et ce sont ces travaux qui vont être examinés en premier. Je me tournerai ensuite vers d'autres résultats – des violations de divers principes normatifs – dont j'essaierai de montrer qu'on peut les expliquer dans le même cadre.

On peut souligner que les résultats qui suivent ne portent pas directement sur le fonctionnement du raisonnement, mais sur l'effet qu'a la prise en compte de raisons sur la prise de décision. Dans une perspective classique sur le raisonnement, on pourrait juger de tels travaux comme n'étant guère pertinents. Cependant, si on accepte l'acception suggérée ici du raisonnement comme mécanisme évaluant des raisons, on voit qu'il est forcément à l'œuvre dès lors que des raisons jouent un rôle dans la prise de décision. De plus, bien que ces travaux ne portent pas directement sur le fonctionnement du raisonnement, on peut tout de même les utiliser pour faire des inférences à ce propos, comme nous allons le voir dans la conclusion.

Il est important de détailler ces travaux et de les passer en revue pour deux raisons. D'une part, il n'est pas forcément aisé de faire la démonstration qu'un choix est bien basé sur des raisons. A cette fin, de nombreuses techniques différentes peuvent être utilisées, et les démonstrations varient d'un cas à l'autre. Il est donc bon de s'attarder sur ces démonstrations pour s'assurer qu'elles sont convaincantes. D'autre part, il est important de montrer qu'il s'agit d'un phénomène très courant, à tel point qu'il explique une part substantielle des biais classiquement étudiés en prise de décision.

8.2.1 **Effets d'attraction et de compromis**

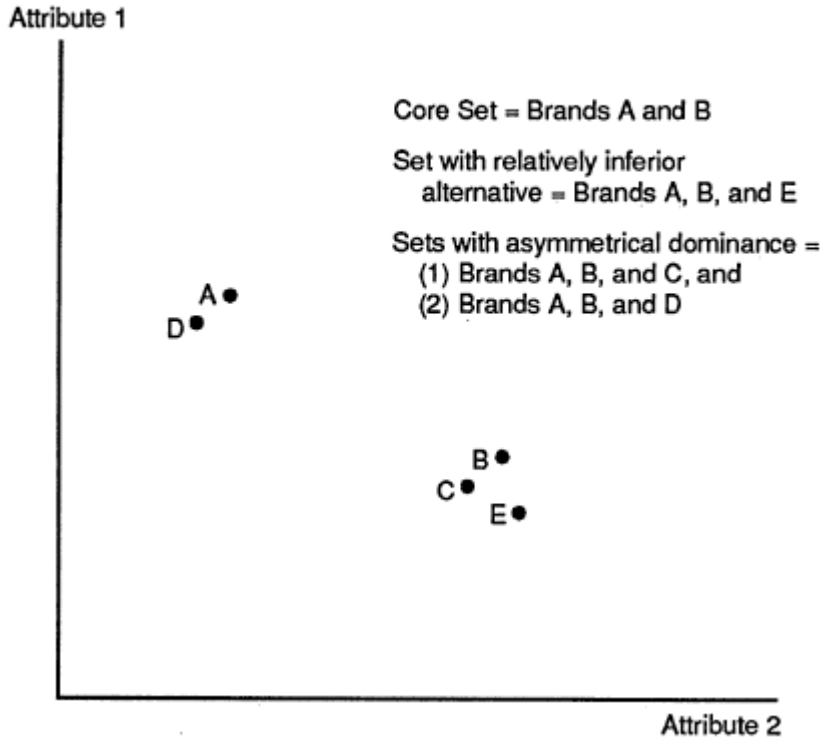
La première étude spécifiquement consacrée au choix basé sur des raisons fut conduite par Itamar Simonson en 1989 (Simonson, 1989). Simonson s'appuie sur quelques travaux précédents en prise de décision ayant mentionné l'idée que certains choix pouvaient s'expliquer par le recours à des *raisons* : « the idea that individual choice behavior under preference uncertainty can be better understood when seen as based on the available reasons or justifications for and against each alternative » (p.158). Il mentionne également de nombreux travaux en psychologie sociale

montrant l'importance de ces justifications dans la vie de tous les jours, et indiquant la possibilité d'une 'intériorisation' de ces justifications. C'est-à-dire que les gens puissent s'en servir pour eux-mêmes, dans des circonstances dans lesquelles il n'est pas clair qu'ils auront à justifier leurs choix. Simonson, cependant, signale également l'aspect post-hoc des explications en termes de raisons qui avaient été avancées jusqu'à présent : elles sont intuitivement justes, mais elles sont toujours avancées après que le résultat à expliquer a été découvert. Simonson va donc chercher à tester des hypothèses plus précises concernant le choix basé sur des raisons. Voici les trois hypothèses qu'il va tester : (i) un choix basé sur des raisons sera renforcé si les participants doivent se justifier ; (ii) un choix basé sur des raisons sera perçu comme étant plus facile à justifier et comme ayant moins de chances d'être critiqué ; (iii) un choix basé sur des raisons devrait donner lieu à des explications plus élaborées. Cette dernière hypothèse découle du fait que les choix basés sur des raisons devraient être moins intuitifs, plus verbaux que les choix pris par d'autres moyens.

Simonson va tester ces hypothèses au moyen de deux effets que l'on peut supposer causés par un choix basé sur des raisons. Le premier est l'effet d'attraction (Huber, Payne, & Puto, 1982). L'effet survient lorsqu'on ajoute à un choix entre deux alternatives de valeur similaire (l'ensemble de base) une alternative qui est dominée (qui est inférieure sur tous les attributs) par une des alternatives initiales mais pas l'autre. On parle alors de dominance asymétrique car un seul des choix initiaux est dominant par rapport à la nouvelle option. L'effet d'attraction peut également survenir si l'alternative ajoutée n'est que légèrement inférieure (et non complètement dominée) par une des alternatives. Dans les deux cas, l'alternative de l'ensemble de base qui est dominante (ou légèrement supérieure) à la nouvelle alternative tendra à être choisie davantage dans ce nouveau contexte, violant ainsi le principe de régularité selon lequel l'ajout d'une alternative ne peut pas augmenter les chances qu'une autre soit choisie (Luce, 1977). La figure suivante (tirée de Simonson, 1989, p.161) illustre l'effet d'attraction :

FIGURE A

THE ATTRACTION EFFECT

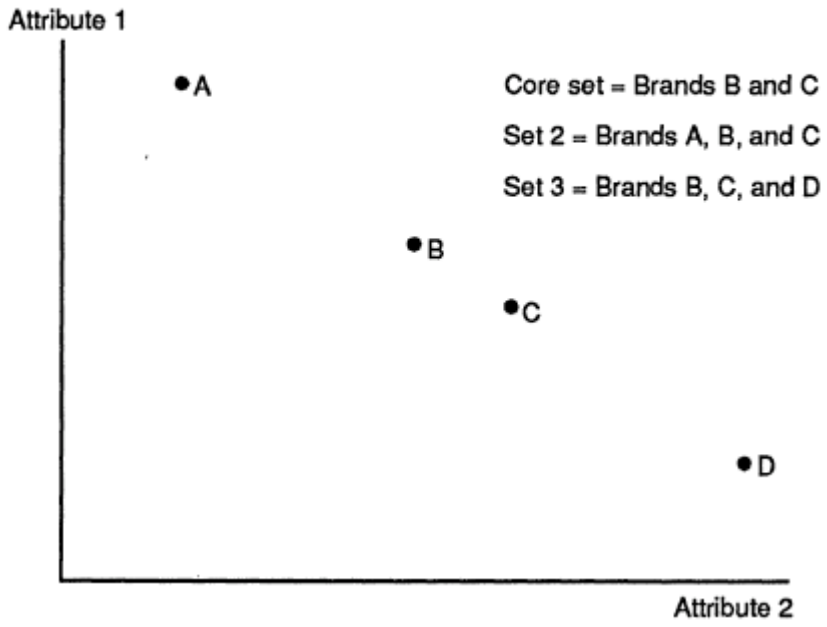


On y voit l'ensemble de base (A et B) et les différentes alternatives qui seront utilisées dans les expériences pour créer un effet d'attraction : D, que A domine et qui devrait renforcer la probabilité qu'il soit choisi, et C (que B domine) et E (qui est légèrement inférieur à B) qui devraient renforcer la probabilité que B soit choisi. L'effet d'attraction peut s'expliquer en termes de raisons car les nouvelles alternatives suggèrent des raisons facilement accessibles pour justifier un choix. Alors que le choix initial entre A et B peut être ardu car les attributs des deux alternatives s'équilibrent, et donc être dur à justifier. Mais l'ajout de ces nouvelles alternatives fournit une solution aisée pour justifier une des réponses.

Le second effet qui est utilisé par Simonson est l'effet de compromis. A nouveau, il s'agit d'une modification non normative de la répartition des choix entre deux alternatives d'un ensemble de base lors de l'ajout d'une nouvelle alternative. Cette fois, l'alternative ajoutée est aussi bonne que les autres, mais son ajout transforme une des alternatives de l'ensemble de base en option de compromis. Ceci est illustré par la figure suivante :

FIGURE B

A COMPROMISE EFFECT BY ADDING A DISTANT COMPETITOR



Il est cependant plus difficile dans ce cas de faire des prédictions sur l'effet de l'ajout de la nouvelle alternative. D'un côté, on peut imaginer qu'être le plus fort sur un des attributs rende une option plus facile à justifier. Dans ce cas, l'ajout de A, par exemple, devrait faire diminuer la part relative de B par rapport à C. D'un autre côté, dans la mesure où les personnes doivent se justifier à un public inconnu, le choix de compromis peut paraître moins risqué : il s'agit de l'option qui est assez bonne quelque soit l'attribut envisagé. On devait alors au contraire observer une augmentation de la part relative de choix B par rapport à C suite à l'introduction de A. Simonson prédit que la manipulation qu'il effectue (et qui sera décrite ci-dessous) aura plutôt pour effet de renforcer l'attrait de l'option qui représente le compromis. Mais on voit que dans ce cas il est possible de trouver des raisons qui aillent dans les deux directions.

L'objectif de la première expérience était de tester l'hypothèse (i) selon laquelle le fait de devoir se justifier devrait renforcer les choix basés sur des raisons. Les participants commençaient par évaluer l'importance de différents attributs (afin de s'assurer que les choix de l'ensemble de base étaient bien équivalents par exemple), puis ils étaient confrontés à une série de choix entre plusieurs objets. Ces choix étaient différents selon les participants, de façon à recouvrir, en inter-participants, les choix représentés dans les figures A, B et C. Finalement l'hypothèse

était testée au moyen de la comparaison entre deux hypothèses. D'un côté une condition 'anonyme', dans laquelle les instructions informaient les sujets que les résultats étaient totalement anonymes. De l'autre une condition 'justification' dans laquelle, au contraire, les participants devaient mettre leur nom sur leurs questionnaires et leurs initiales sur chaque page, afin que leurs réponses puissent être utilisées comme exemple en classe. Ils pourraient alors être invités à justifier leurs choix.

Les résultats furent globalement conformes aux prédictions. Tout d'abord, les effets d'attraction et de compromis furent bien observés. La manipulation de justification eut un effet fort (et significatif) sur l'effet d'attraction. Par exemple, pour un ensemble de choix (voir figure A), l'effet total d'attraction[63] était plus important de 17% dans la condition 'justification'. Les résultats positifs dans le cas de l'effet d'attraction furent également observés dans le cas de l'effet de compromis, à quelques nuances peu pertinentes près.

La deuxième expérience visait, elle, à tester l'hypothèse (ii) selon laquelle les choix basés sur des raisons devraient être plus faciles à justifier et avoir moins de chances d'être critiqués. Pour cela, un nouvel ensemble de participants dut remplir trois tâches. La première tâche consistait à évaluer exactement les mêmes choix que ceux qui avaient été présentés aux participants de l'expérience 1. Cette fois cependant ce n'était pas les objets eux-mêmes qui devaient être évalués, mais les chances qu'ils avaient d'être critiqués et, à l'inverse, la facilité avec laquelle ils pourraient être justifiés. La seconde tâche était similaire à la première, mais cette fois les objets étaient évalués en dehors du contexte qui en faisait de bons choix basés sur des raisons. Ils étaient donc évalués pour leurs attributs mêmes. Les participants devaient alors choisir quelle alternative avait le plus de chances d'être choisie par un étudiant pour qui les critiques potentielles, et la possibilité de se justifier sont très importantes. Finalement, ils devaient se prononcer sur des questions spécifiquement liées à l'effet de compromis : critiqueraient-ils plus facilement un étudiant faisant un choix de compromis ou un autre choix ?

Concernant l'effet d'attraction, les résultats confirmèrent les prédictions : les choix basés sur des raisons étaient vus comme étant moins facilement critiquables et

[63] Ici calculé comme la différence entre l'augmentation de l'option A lorsqu'elle est dominante et sa diminution lorsque B est dominante.

plus faciles à justifier. Le cas de l'effet de compromis est plus ambigu. Si l'option de compromis était vue comme plus dure à critiquer, elle n'était pas perçue comme étant plus facilement justifiable. De plus, les opinions vis-à-vis des personnes choisissant l'option de compromis étaient partagées. Deux tiers des participants déclarèrent avoir plus de chances de critiquer un choix autre que le compromis. Mais d'autres personnes jugèrent au contraire qu'un tel choix pouvait au contraire révéler une trop grande envie de plaire, l'absence d'un vrai choix personnel, etc. On retrouve donc ici l'intuition mentionnée plus haut que la transformation d'une option en compromis pouvait jouer en sa faveur aussi bien qu'en sa défaveur – selon les objets, les juges, etc.

Finalement, l'expérience 3 fut conduite afin de tester l'hypothèse (iii) selon laquelle les choix basés sur des raisons seraient assortis d'explications plus élaborées. Pour cela, l'auteur eut recours à un protocole de pensée à haute voix, les pensées étant récoltées alors que de nouveaux participants répondaient aux questions utilisées dans l'expérience 1. Pour ce qui est de l'effet d'attraction, il convient tout d'abord de distinguer deux types de choix. Lorsqu'un participant choisit l'option dominante, il peut le faire pour deux raisons principales. La première est que cette option est supérieure aux deux autres options sur l'attribut que ce participant juge comme étant le plus important (choix lexicographique). Dans ce cas, l'introduction de l'option dominée ne joue aucun rôle. Mais il peut également choisir cette option à cause de l'effet d'attraction lui-même : car l'option dominée fournit des raisons de choisir l'option dominante. Etant donné que les participants devaient indiquer, avant de remplir la tâche de choix elle-même, l'importance des différents attributs à leurs yeux, il est possible de comparer les pensées exprimées dans les deux cas. La prédiction est alors que les pensées des participants devraient être plus élaborées dans le cas du choix basé sur des raisons. C'est bien ce qui fut observé. Ces protocoles étaient plus significativement plus longs, ils faisaient plus de référence à la difficulté de la tâche et mentionnaient plus souvent les avantages et désavantages de l'option sélectionnée (toutes différences significatives). Enfin, il est intéressant de noter que ces choix basés sur des raisons se faisaient souvent au détriment de la valeur intrinsèque des options : parmi les participants choisissant l'option dominante, plus de 60% le firent alors même que l'importance des attributs qu'ils avaient eux-mêmes évalués aurait dû les mener à choisir l'autre option de l'ensemble de base. Les résultats furent globalement similaires dans le cas de l'effet de compromis. Les

protocoles des participants choisissant les options de compromis étaient plus longs, faisaient plus de référence à la difficulté de la tâche et aux avantages et inconvénients de l'option choisie. De plus, de nombreux participants eurent directement recours à des explications du type 'cette option représente un compromis'.

Ces expériences sont intéressantes par de nombreux aspects. Tout d'abord, il s'agit de la première démonstration de choix basé sur des raisons. D'autres résultats – dont certains seront passés en revue plus bas – avaient déjà reçu une explication dans ce cadre, mais elle n'avait jamais été réellement démontrée. Ici, plusieurs éléments indiquent que les participants font certains choix car il leur est plus facile de les justifier au moyen de certaines raisons : la fréquence de ces choix augmente avec le besoin de se justifier, ils sont jugés par d'autres comme étant moins critiquables et plus faciles à justifier, et les participants raisonnent plus, fournissent plus d'explications lorsqu'ils les prennent. Dans ces expériences, le raisonnement a des conséquences négatives. A cause de lui, les participants tendent à choisir des options qui ne sont pas conformes à leurs propres critères. De plus, l'analyse des protocoles montre que ces erreurs tendent à être commises davantage par les gens qui raisonnent le plus, qui pèsent le pour et le contre, qui font donc ce qui est habituellement recommandé pour prendre de bonnes décisions.

Finalement, il est important de souligner que ces choix ne sont pas faits uniquement lorsque les participants pensent qu'ils devront justifier leur réponse, mais également dans des conditions d'anonymat. On pourrait imaginer que les causes du même comportement non normatif soient différentes dans les deux conditions, mais cette explication serait bien peu économique, et il faudrait alors une autre explication pour les effets d'attraction et de compromis. Ces résultats signifient que les gens ont bien tendance, dans ce type de situation, à chercher des raisons, à justifier leurs choix, et ce même lorsque le contexte ne les y invite pas particulièrement – lorsqu'il s'agit bien, en principe tout du moins, de prise de décision individuelle. Il est cependant intéressant de constater que dans ces expériences une autre condition qui joue un rôle pour le déclenchement du raisonnement dans la théorie argumentative est bien remplie. Il s'agit de la faiblesse des intuitions. En effet, les participants doivent choisir entre des objets décrits de façon très abstraite (un pack de bière coûtant 1$90 et étant noté 65 en qualité ou un pack de bière coûtant 2$80 et étant noté 75 ?). Le lien entre la faiblesse des intuitions et l'utilisation de raisons est

indiqué par le fait que les participants ayant fait les choix basés sur des raisons rapportèrent beaucoup plus souvent que la tâche était difficile, et donc qu'aucune solution intuitive ne s'imposait d'elle-même. Cette interprétation est renforcée par une série d'expériences ayant montré que plus les participants étaient familiers avec les produits, ou plus leurs descriptions était élaborées, moins l'effet d'attraction était fort (Ratneshwar, Shocker, & Stewart, 1987) (voir Chernev, 2005, pour des expériences similaires concernant l'effet des attributs qui s'équilibrent bien).

Simonson a également conduit d'autres études sur l'effet de compromis et le choix basé sur des raisons. Une d'entre elle concerne le rôle d'une variable de personnalité (le besoin d'unicité, Simonson & Nowlis, 2000), mais elle sera examinée plus bas car elle ne contient qu'une expérience portant spécifiquement sur l'effet de compromis parmi de nombreuses autres. Une autre s'est penchée sur les variations interculturelles (Briley, Morris, & Simonson, 2000). Elle teste l'hypothèse que les raisons les plus accessibles, ou celles qui sont perçues comme étant les plus persuasives, varient selon les cultures. Plus spécifiquement, les raisons amenant les gens à choisir les options de compromis peuvent être plus ou moins accessibles ou persuasives. Les cultures qui sont comparées sont une culture occidentale (Etats-Unis d'Amérique) et des cultures orientales (Japon et Hong-Kong). Sur la base de divers travaux interculturels, les auteurs prédisent que les orientaux devraient avoir recours plus facilement aux raisons poussant vers l'option de compromis que les occidentaux.

Les expériences de Simonson (1989) ont montré que le besoin de se justifier pouvait augmenter, chez des occidentaux, la tendance à choisir l'option de compromis. Il faut souligner que dans ces expériences, les participants de la condition 'justification' pensaient être évalués et devoir justifier leurs choix devant leur classe. Il s'agit donc de circonstances dans lesquelles la prudence est de mise : il s'agit de choisir une option qui a peu de chances d'être critiquée au moins autant qu'une option qu'on peut justifier. Ces circonstances sont particulièrement propices au choix de l'option de compromis. Cette option a moins de chances d'être critiquée et, face à un public dont on ne connaît pas les préférences, on a plus de chances de pouvoir la justifier. Cela ne veut cependant pas dire qu'il n'est pas au moins aussi facile de trouver des raisons pour les options 'extrêmes' : on peut facilement imaginer arguer qu'un attribut est plus important que l'autre par exemple. La

supériorité de l'option de compromis dans la condition justification peut donc être due à un contexte évaluatif assez particulier. Il est fort possible que dans d'autres contextes – en particulier si les participants doivent uniquement donner des raisons, sans avoir peur de se faire juger – l'effet soit en fait inversé. C'est d'ailleurs ce qu'ont observé Simonson et Nowlis (2000) dans une expérience sur laquelle nous reviendrons. Si, donc, l'option de compromis pourrait se trouver désavantagée par le fait de devoir fournir des raisons chez des participants occidentaux, Briley et al. font au contraire l'hypothèse que dans certaines cultures orientales l'option de compromis sera tout de même choisie davantage dans ces contextes. Elle pourrait l'être parce que la peur de la critique y est plus omniprésente, ou que les justifications pour cette option y sont plus disponibles ou persuasives.

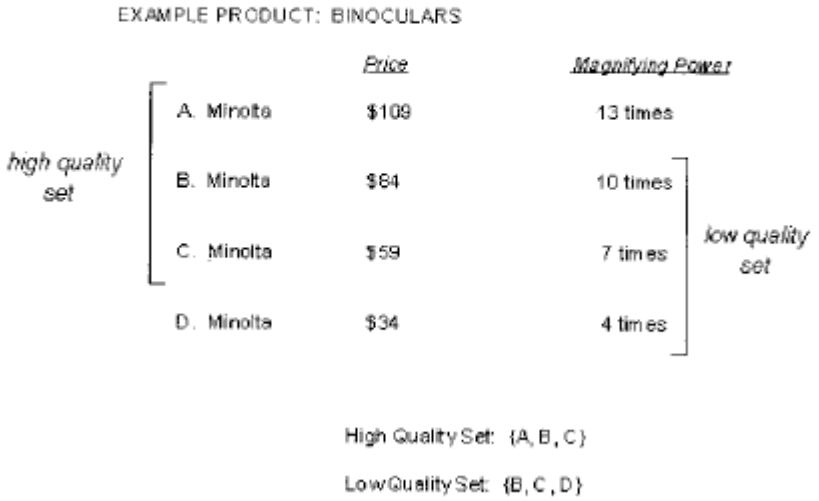
Afin de tester cette hypothèse, les auteurs conduisirent cinq expériences. Dans la première, des participants d'une université américaine et d'une université hongkongaise devaient faire des choix parmi des triplets d'objets caractérisés par deux attributs. Les attributs étaient ainsi répartis qu'il y avait deux options extrêmes alors que la dernière représentait un compromis. Par exemple, ils pouvaient avoir à choisir entre un ordinateur avec 48 mégaoctets de RAM et un giga-octet de disque dur, un ordinateur avec 16 mégaoctets de RAM et trois giga-octets de disque dur et un ordinateur (le compromis) avec 32 mégaoctets de RAM et deux gigaoctet de disque dur. Dans une condition, les participants devaient donner les raisons pour leur choix. Deux aspects de ces expériences sont à souligner. D'une part les raisons devaient être fournies avant d'indiquer le choix, augmentant ainsi les chances qu'elles exercent une influence. D'autre part, il ne s'agit pas d'un contexte aussi fortement évaluatif que dans l'expérience de Simonson (1989). Bien que les participants aient conscience qu'une personne va lire et analyser les raisons qu'ils notent, le questionnaire reste anonyme et ils n'auront pas à justifier eux-mêmes leurs choix devant leur classe par exemple.

Les prédictions furent confirmées par les résultats. Alors qu'aux Etats-Unis le fait de devoir donner des raisons augmentait le pourcentage de réponses extrêmes (de 52% à 61%), à Hong-Kong ce sont les choix de compromis qui augmentaient (de 49% à 56%), les deux différences étant significatives. Ces différences étaient visibles dans le type de raison qui tendait à être fourni par les participants des deux cultures. Ces résultats furent répliqués dans une seconde expérience comparant cette fois des participants Japonais aux participants Américains.

Dans une nouvelle expérience, Briley et al. eurent recours à des participants d'origine asiatique mais ayant passé plusieurs années aux Etats-Unis, les comparant à des participants natifs des Etats-Unis et d'ascendance européenne. Cette fois, l'effet de compromis fut mesuré en comparant (entre participants) les choix entre deux ensembles tels quel ceux représentés dans la figure suivante (tiré de Briley et al., 2000, p.167) :

FIGURE 4

STUDY 3 CHOICE SET CONFIGURATION



L'effet de compromis s'observe par un choix plus important de B par rapport à C dans le groupe de gauche (dans lequel B est l'option de compromis) par rapport au groupe de droite (dans lequel B est une option extrême). Les résultats indiquèrent un effet du fait de devoir fournir des raisons chez les participants Européens-Américains. Chez eux, le choix de B fut renforcé lorsqu'il était en position extrême. Il n'y eut par contre pas d'effet chez les participants d'origine asiatique, dont on peut justement imaginer qu'ils étaient à un stade d'assimilation dans lequel des influences des deux cultures pourraient se faire sentir.

La quatrième expérience portait sur une source potentielle de la préférence pour les raisons soutenant l'option de compromis chez les orientaux : les proverbes. De par leur ancienneté, ceux-ci sont censés refléter des différences persistantes entre cultures. Des recueils de proverbes Chinois et Américains furent tout d'abord examinés par des juges pour déterminer s'ils promouvaient les solutions de compromis ou au contraire les solutions extrêmes (ou ni l'une ni l'autre de ces options), avant d'être évalués par des participants Américains et Hongkongais. Le

premier résultat était une plus grande proportion de proverbes en faveur des options de compromis parmi les proverbes Chinois. Le second que les participants Américains avaient une préférence plus marquée pour les proverbes défendant des options extrêmes que les Hongkongais, et vice versa pour ceux se faisant l'avocat du compromis.

Finalement, dans la cinquième expérience les participants devaient évaluer les raisons données par un autre participant : ils devaient déterminer si ces raisons étaient persuasives. Ces raisons pouvaient favoriser le compromis (« The average option is usually a good one, so I choose the one in the middle ») ou non (« It is important to figure out exactly what I want and not to settle for something average »). Alors que les participants de Hong-Kong notèrent les deux types de raisons également, ceux des Etats-Unis accordèrent une préférence significative aux raisons favorisant une solution extrême.

Tous ces résultats renforcent la théorie du choix basé sur des raisons. Il est clair que dans ce cas, l'accessibilité de différentes raisons joue un rôle causal dans les décisions prises par les participants. L'extension au domaine interculturel est également intéressante. Elle montre la variabilité des raisons qui sont préférées dans différentes cultures, et offre ainsi un moyen d'expliquer certaines différences interculturelles sans faire appel à des différences psychologiques profondes.

8.2.2 Effet de disjonction

L'effet de disjonction est une autre violation des normes de la théorie de la décision pour laquelle des explications en termes de choix basé sur des raisons ont été avancées. Il s'agit d'une violation du principe de la chose sûre. Ce principe dit que si une personne préfère A à B si un certain événement intervient, et préfère également A à B si ce même événement n'intervient pas, alors elle devrait préférer A à B quelque soit le degré d'incertitude vis-à-vis de la survenue de l'événement en question (Savage, 1954). Dans une série d'articles, Tversky et Shafir (Shafir & Tversky, 1992; Tversky & Shafir, 1992b) ont démontré plusieurs violations de ce

principe, comme dans l'exemple suivant (il s'agit du pourcentage de participants ayant donné chaque réponse)[64] :

> Disjunctive version:
> Imagine that you have just taken a tough qualifying examination. It is the end of the fall quarter, you feel tired and run-down, and you are not sure that you passed the exam. In case you failed you have to take the exam again in a couple of months - after the Christmas holidays. You now have an opportunity to buy a very attractive 5-day Christmas vacation package in Hawaii at an exceptionally low price. The special offer expires tomorrow, while the exam grade will not be available until the following day. Would you?:
>
> | | |
> |---|---|
> | (a) buy the vacation package. | 32% |
> | (b) not buy the vacation package. | 7% |
> | (c) pay a $5 non-refundable fee in order to retain the rights to buy the vacation package at the same exceptional price the day after tomorrow - after you find out whether or not you passed the exam. | 61% |
>
> Pass / fail versions:
> Imagine that you have just taken a tough qualifying examination. It is the end of the fall quarter, you feel tired and run-down, and you find out that you [passed the exam./failed the exam. You will have to take it again in a couple of months - after the Christmas holidays.] You now have an opportunity to buy a very attractive 5-day Christmas vacation package in Hawaii at an exceptionally low price. The special offer expires tomorrow. Would you?:
>
> | | Pass | / Fail |
> |---|---|---|
> | (a) buy the vacation package. | 54% | 57% |
> | (b) not buy the vacation package. | 16% | 12% |

[64] L'explication qui suit s'inspire de celle donnée dans Shafir et al. (1993).

(c) pay a $5 non-refundable fee in order to retain the rights to buy the vacation package at the same exceptional price the day after tomorrow.	30%	31%

On voit que plus de participants choisirent d'acheter les vacances à Hawaï lorsqu'ils connaissaient le résultat de l'examen que lorsqu'ils ne le connaissaient pas et ce *quelque soit le résultat de l'examen*. Pour les auteurs, ce phénomène s'explique par la relative disponibilité de raisons pour les différents choix. Lorsque les participants pensent qu'ils ont ou n'ont pas eu l'examen, ils ont une bonne raison pour prendre des vacances dans les deux cas (comme récompense, ou comme façon de se changer les idées après cette déception par exemple). Ces raisons cependant, bien que pointant dans la même direction, sont différentes et même, dans une certaine mesure, incompatibles. Donc les participants qui ne savent pas s'ils ont eu l'examen ou non ont du mal à se décider : les éventuelles raisons pouvant les décider à partir sont contradictoires.

Bien que les auteurs n'aient pas fait de manipulations visant à confirmer cette explication de la même façon que l'avait fait Simonson ils avancent le résultat suivant pour montrer que l'effet n'est pas dû à la disjonction elle-même, mais bien à la disjonction entre les raisons avancées pour le même choix. Il s'agit d'une comparaison entre deux situations. D'un côté ces situations sont structurellement similaires car elles impliquent toutes les deux la même incertitude vis-à-vis d'un événement. D'un autre côté elles sont distinctes car dans un cas les raisons pour faire le choix sont différentes selon que l'événement survient ou non alors que dans un autre cas elles sont identiques.

La première situation est la suivante :

Win / lose version:

Imagine that you have just played a game of chance that gave you a 50% chance to win $200 and a 50% chance to lose $100. The coin was tossed and you have [won $200/lost $100. You are now offered a second identical gamble: 50% chance to win $200 and 50% chance to lose $100. Would you?:

	Won	Lost
(a) accept the second gamble.	69%	59%
(b) reject the second gamble.	31%	41%

Disjunctive version:

Imagine that you have just played a game of chance that gave you a 50% chance to win $200 and a 50% chance to lose $100. Imagine that the coin has already been tossed, but that you will not know whether you have won $200 or lost $100 until you make your decision concerning a second, identical gamble: 50% chance to win $200 and 50% chance to lose $100. Would you?:

(a) accept the second gamble.	36%
(b) reject the second gamble.	64%

Il s'agit là d'une réplication de l'effet de disjonction tel qu'il avait été démontré dans l'expérience précédente : plus de personnes acceptent le second pari lorsqu'ils connaissent le résultat du premier (quel qu'il soit) que lorsqu'ils ne le connaissent pas. Selon les auteurs, les raisons de l'acceptation du second pari sont différentes selon le résultat du premier. Si le premier est gagné, alors les participants peuvent se dire qu'ils ont déjà gagné 200$, il est facile d'en risquer la moitié pour un pari qui a des chances de leur en rapporter encore plus. Si le premier pari est perdu, ils peuvent voir le second comme une « chance de sortir du rouge » (Shafir et al., 1993, p.29). Ces deux raisons ne sont pas compatibles, ce qui causerait l'effet de disjonction observé ici. Les auteurs contrastent ce cas avec une situation analogue dans laquelle les raisons pour choisir le second pari devraient être les mêmes quelque soit le résultat du premier pari :

Imagine that you have just played a game of chance that gave you a 50% chance to win $600 and a 50% chance to win $300. Imagine that the coin has already been tossed, but that you will not know whether you have won $600 or $300 until you make your decision concerning a second gamble: 50% chance to win $200 and 50% chance to lose $100.

Dans ce cas, le premier pari, quelque soit son résultat, garanti un gain aux participants. Ils devraient alors accepter le second pari (qui est le même que dans la situation précédente), dans tous les cas, qu'ils sachent qu'ils ont gagné 300$, ou 600$, ou qu'ils ne sachent pas combien ils ont gagné. C'est bien ce qui fut observé : les pourcentages d'acceptation du second pari étaient, respectivement, 69%, 75% et

73%. Les pourcentages sont un peu plus élevés que dans le problème précédent, probablement car les gains plus importants renforcent la disponibilité de la raison en termes de gains déjà acquis, mais le plus intéressant est la disparition totale de l'effet de disjonction, ainsi que prévu.

Shafir et Tversky (1992) ont également utilisé cet effet de disjonction résultant d'un choix basé sur des raisons pour expliquer des comportements relativement étranges face au dilemme du prisonnier. Le dilemme du prisonnier met le participant dans la peau d'un détenu qui a été capturé avec son complice. La situation à laquelle il fait face est la suivante. S'il ne dénonce pas son complice, et que son complice ne le dénonce pas non plus, alors ils s'en tirent tous les deux avec une peine légère ; par contre, si son compère le trahit, il encourra la peine maximale. S'il trahit son compère et que celui-ci le trahi également, ils encourent tous les deux une peine moyenne (étant récompensés pour leur coopération avec la police), mais si ce dernier ne le trahi pas, alors le participant est libre. Dans le cas de l'expérience conduite par Shafir et Tversky, la pondération précise des différentes options était la suivante (Shafir et Tversky, 1992, p.452) :

		OTHER	
		cooperates	competes
YOU	cooperate	You: 75 Other: 75	You: 25 Other: 85
	compete	You: 85 Other: 25	You: 30 Other: 30

Les nombres représentent les gains pour les deux personnes. Pour poursuivre l'analogie, on peut dire que plus ils sont élevés, plus la peine de prison est courte (il faut préciser que 'coopérer' et 'être en compétition' s'appliquent ici l'action vis-à-vis de son complice et non de la police). Deux éléments sont capitaux et constituent la

particularité du dilemme du prisonnier. D'une part le gain total maximum, pour les deux joueurs, est atteint s'ils coopèrent tous les deux. Cependant, la meilleure stratégie est toujours d'être en compétition. En effet, si l'autre coopère et que le participant ne coopère pas, il gagne 85 à la place de 75 ; si l'autre ne coopère pas et que le participant ne coopère pas non plus, il gagne 30 à la place de 25. Etre en compétition est donc la stratégie dominante dans ce jeu, bien qu'elle soit contre-productive si adoptée par les deux joueurs qui alors auraient mieux fait de coopérer tous les deux (gagnant chacun 75 plutôt que 30 – on dit alors que l'équilibre dû à la stratégie dominante n'est pas Pareto optimal). De très nombreux résultats indiquent cependant qu'une proportion considérable de personnes choisit la coopération. Shafir et Tversky proposent une analyse de ce problème comme un choix basé sur des raisons.

Plus précisément, ils font l'hypothèse que lorsque le participant ne sait pas encore ce que va faire l'autre joueur, il est assez facile de trouver des raisons pour coopérer : les gains potentiels de la coopération semblent importants. Par contre, si le joueur sait ce que l'autre a fait, alors il devient facile de trouver des raisons pour être en compétition : il suffit de comparer les deux options, et on se rend compte qu'être en compétition est supérieur, que l'autre ait coopéré ou non. Afin de tester ces hypothèses, ils soumirent à des participants plusieurs dilemmes du prisonnier : certains sous leur forme normale (dans laquelle les participants ne savent pas ce que l'autre joueur a fait), et d'autres sous une forme modifiée dans laquelle les participants étaient informés que l'autre jouer avait coopéré ou non. Les résultats furent conformes aux prédictions. Lorsque l'autre joueur n'avait pas coopéré, 3% des participants coopérèrent ; lorsque l'autre avait coopéré, ils étaient 16% à le faire également ; mais ils étaient 37% à le faire lorsqu'ils ne savaient pas quelle stratégie l'autre avait adoptée (toutes différences significatives). Cela signifie que plus de candidats sont prêts à coopérer quand l'autre peut soit coopérer soit les trahir que lorsqu'ils savent que l'autre a coopéré. Il s'agit même là du second pattern le plus courant (derrière être en compétition dans toutes les circonstances), représentant près d'un tiers des participants.

Plus récemment, Croson (1999) a répliqué ce résultat en utilisant une méthodologie quelque peu différente, et l'a étendu à d'autres jeux (dilemme du prisonnier asymétrique, jeux de dominance itérée). De plus, il a mis au point dans le cas du dilemme du prisonnier une expérience similaire à celle présentée plus haut

dans le cas des paris. Il a comparé un dilemme du prisonnier classique à une version aussi complexe mais pour laquelle les raisons en faveur des choix initiaux étaient toujours les mêmes. Dans ce cas, l'effet de disjonction disparut : les participants faisaient les mêmes choix dans toutes les situations. Ce résultat confirme donc que l'effet de disjonction observé dans le dilemme du prisonnier résulte bien du fait que les choix sont guidés par des raisons différentes selon les circonstances.

8.2.3 Autres effets du choix basé sur des raisons

D'autres phénomènes liés à celui qui vient d'être décrit peuvent s'expliquer en termes de choix basé sur des raisons. Ils ne seront examinés ici que rapidement car, bien que leur explication en termes de choix basé sur des raisons soit très plausible, elle n'a pas été précisément démontrée, par exemple au moyen des techniques utilisées par Simonson (1989).

Nous venons de voir que l'effet de disjonction était causé par la présence de raisons contradictoires. Dans d'autres cas, des raisons peuvent s'annuler mutuellement car elles sont trop similaires : si on a une bonne raison de faire un choix A, et une raison similaire et aussi forte de faire un choix B, alors on n'a pas de bonne raison de choisir A plus que B ou vice versa. Cet effet a été observé dans les situations suivantes (Tversky & Shafir, 1992a; Tversky & Simonson, 1993) :

> High conflict:
> Suppose you are considering buying a compact disk (CD) player, and have not yet decided what model to buy. You pass by a store that is having a l-day clearance sale. They offer a popular SONY player for just $99, and a top-of-the-line AIWA player for just $169, both well below the list price. Do you?:
>
> | | |
> |---|---|
> | (x) buy the AIWA player. | 27% |
> | (y) buy the SONY player. | 27% |
> | (z) wait until you learn more about the various models. | 46% |

Low conflict:

Suppose you are considering buying a CD player, and have not yet decided what model to buy. You pass by a store that is having a l-day clearance sale. They offer a popular SONY player for just $99, well below the list price. Do you?:

(y) buy the SONY player.	66%
(z) wait until you learn more about the various models.	34%

Les résultats montrent que les participants ont plus de chances d'acheter un lecteur CD (plutôt que de remettre leur choix à plus tard) lorsqu'ils ont plus de choix. En effet, les raisons pour choisir le lecteur SONY ou l'AIWA sont similaires (les deux sont en soldes), et elles ne permettent pas de les départager. Les participants sont donc plus motivés pour attendre et trouver une autre raison. Que cet effet ne soit pas dû simplement à la présence d'un choix supplémentaire dans la situation de conflit est démontré par la condition suivante :

Dominance:

Suppose you are considering buying a CD player, and have not yet decided what model to buy. You pass by a store that is having a l-day clearance sale. They offer a popular SONY player for just $99, well below the list price, and an inferior AIWA player for the regular list price of $105. Do you?:

(x') buy the AIWA player.	3%
(y) buy the SONY player.	73%
(z) wait until you learn more about the various models.	24%

On voit que dans ce cas, et bien qu'il y ait la même quantité d'items, les participants ne choisissent pas de remettre le choix à plus tard : ils ont une bonne raison de choisir un lecteur CD plutôt que l'autre, et n'ont donc pas besoin d'attendre d'avoir des raisons supplémentaires.

Un autre phénomène lié à la fois à celui qui vient d'être revu et à l'effet d'attraction dû à une dominance asymétrique observé par Simonson (1989) peut s'observer si on compare les deux situations suivantes (Tversky et Shafir, 1992) :

> Conflict:
> Imagine that you are offered a choice between the following two gambles:
> (x) 65% chance to win $15
> (y) 30% chance to win $35
>
> Dominance:
> Imagine that you are offered a choice between the following two gambles:
> (x) 65% chance to win $15
> (x') 65% chance to win $14
>
> In both cases: You can either select one of these gambles or you can pay $1 to add one more gamble to the choice set. The added gamble will be selected at random from the list you reviewed [les participants avaient précédemment vu une liste de 12 paris dont les gains anticipés étaient similaires à ceux des paris presentés].

Dans ce cas, les participants furent 64% à demander l'alternative supplémentaire dans la condition 'conflit' mais seulement 32% à le faire dans la condition 'dominance' (une différence significative à travers plusieurs problèmes de ce type). En termes de choix basé sur des raisons, cela s'explique par l'absence de raisons claires de choisir un pari plutôt que l'autre dans la situation de conflit, alors qu'au contraire une telle raison est facilement accessible lorsqu'un des paris domine l'autre. Ce comportement ne pourrait pas s'expliquer si les choix étaient basés sur l'attribution d'une valeur à chaque pari : dans ce cas, les participants devraient choisir l'option d'attendre moins souvent (ou au moins aussi peu souvent) dans la condition de conflit que dans la condition de dominance (car le pari qui est choisi dans la condition de dominance est également présent dans la condition de conflit).

Un autre effet qu'on peut facilement expliquer dans le cadre du choix basé sur des raisons est lié au type de raison qui est rendu plus satisfaisant par différentes

questions. Shafir (1993) a ainsi fait l'hypothèse que les réponses aux questions en termes d'acceptation devraient être plus faciles à justifier à l'aide de traits positifs alors qu'à l'inverse les réponses à des questions de termes de rejet devraient être plus faciles à justifier à l'aide de caractéristiques négatives. Afin de tester cette hypothèse, il a croisé deux paramètres : le type de question posé (acceptation ou rejet) et la présence de nombreuses caractéristiques positives et négatives. Alors qu'une option est moyenne sur tous les critères, l'autre est très polarisée : très bonne sur certains critères, très mauvaise sur d'autres. Cette option devrait être à la fois plus facile à choisir car les gens utilisent alors facilement les critères positifs comme raisons pour justifier leur choix, mais aussi et plus facile à rejeter, car ils ont alors recours aux critères négatifs pour justifier leur rejet. C'est ce qui fut observé à travers de nombreux problèmes, dont le suivant est un exemple :

> Imagine that you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations. [To which parent would you award sole custody of the child?/Which parent would you deny sole custody of the child?]

	Award	Deny
Parent A:		
average income		
average health		
average working hours		
reasonable rapport with the child		
relatively stable social life	36%	45%
Parent B:		
above-average income		
very close relationship with the child		
extremely active social life		
lots of work-related travel		
minor health problems	64%	55%

On voit que le parent A est ici l'option moyenne et le parent B l'option extrême. Les résultats sont bien conformes aux prédictions : quelque soit la question posée, c'est le parent B qui est choisi – ce qui signifie qu'il est à la fois plus souvent choisi *et* plus souvent rejeté.

8.2.4 Effets de raisons non pertinentes

Parmi les raisons qui peuvent pousser les gens à faire certains choix, certaines ne sont clairement pas pertinentes. On peut dans ce cas se demander pourquoi elles sont prises en compte, et quels seront leurs effets. Dans une première série d'études, Simonson et ses collègues ont démontré que des raisons n'ayant manifestement aucune pertinence pour les participants pouvaient les pousser à rejeter les choix associés. Par exemple, les participants apprenant qu'une autre personne avait choisi un MBA pour une raison qui lui était propre (avoir de la famille habitant à côté de l'université proposant le MBA) avaient moins de chances de le choisir que des participants à qui aucune information supplémentaire n'était donnée (Simonson, Nowlis, & Simonson, 1993). Si on voit bien pourquoi la raison ne devrait pas pousser les participants à faire le même choix que l'autre personne, le fait de refuser cette alternative n'est pas logique pour autant. De même, Simonson et d'autres collègues ont montré que les attributs promotionnels manifestement inutiles, telle qu'une réduction du prix d'une assiette de collection pour l'achat d'un paquet de céréales, pouvaient fournir des raisons de ne pas choisir les items qui en étaient assortis (Simonson, Carmon, & O'Curry, 1994).

Plus récemment, d'autres auteurs ont cherché à élargir les explications concernant les attributs non pertinents (Brown & Carpenter, 2000). En effet, depuis les résultats obtenus par Simonson, d'autres études avaient donné des résultats inverses. Dans certains cas, des attributs non pertinents pouvaient augmenter les choix associés (voir par exemple Carpenter, Glazer, & Nakamoto, 1994). Les auteurs de cette nouvelle étude utilisent également le cadre du choix basé sur des raisons pour réconcilier ces résultats apparemment contradictoires. Ils s'intéressent aux attributs non pertinents tels que le fait de rajouter de la soie dans les shampoings : il était connu des participants que cette addition ne modifie en rien les propriétés pertinentes du shampoing. Selon les auteurs, ce type d'attribut n'est pas associé à des

raisons de valence positive ou négative en mémoire (au contraire par exemple d'attributs comme 'rend les cheveux plus doux'), et il pourra donc être utilisé, selon les circonstances, comme une raison pour choisir ou pour rejeter une option. L'utilisation qui en est faite dépend alors de la façon dont les raisons peuvent servir l'objectif global qui est de faire un choix (en fait, on devrait plutôt dire de faire un choix justifiable, car s'il s'agissait juste de faire un bon choix, l'analyse suivante ne fonctionnerait pas). Si on doit choisir entre deux items dont un a un attribut non pertinent, alors cet attribut peut être utilisé soit comme raison de choisir, soit comme raison de rejeter l'item en question. Si par contre on doit faire un choix entre trois items, alors la valence des raisons recherchées dépendra du nombre d'items possédant l'attribut non pertinent : si un seul item le possède, alors on peut le considérer comme une raison positive qui nous permet de choisir cet item ; mais si deux items l'ont, alors il est plus simple de le considérer comme une raison de les rejeter car cela permet de choisir le dernier item. Les raisons associées à la présence d'un attribut non pertinent devraient donc dépendre de la situation dans laquelle il est présenté, ce qui a son tour devrait influencer le choix.

La première expérience visait à répliquer certains résultats précédents tout en les comparant afin de tester l'explication avancée. Deux facteurs étaient manipulés : la taille de l'ensemble parmi lequel les participants devaient faire le choix, et la présence ou l'absence d'un attribut non pertinent pour un des items. Rappelons les prédictions : la présence d'un attribut non pertinent devrait augmenter le choix de l'item lorsqu'il est opposé à deux autres items, mais devrait avoir des conséquences mitigées (plutôt négatives à en croire les résultats précédents de Simonson et al., 1994) lorsqu'il est opposé à un seul item. C'est bien ce qui fut observé. Les participants confrontés à un choix parmi trois objets dont un seul avait un attribut non pertinent tendaient à le sélectionner davantage que lorsqu'il en était dépourvu (bien que la différence ne soit que marginalement significative). Dans le cas des ensembles de deux objets, la tendance était inverse (mais même pas marginalement significative).

La seconde expérience visait à tester plus avant l'idée que les attributs non pertinents pouvaient être utilisés comme des raisons positives ou négatives, ainsi qu'à confirmer certains résultats de la première expérience. Cette fois, tous les ensembles comprenaient trois objets. Une comparaison implique un ensemble dans lequel aucun objet n'a d'attribut non pertinent et un ensemble dans lequel seul un

objet en possède un. On s'attend alors à ce que l'attribut soit utilisé comme une raison pour choisir l'objet. L'autre comparaison implique un ensemble dans lequel tous les objets possèdent l'attribut non pertinent et un ensemble dans lequel seul un objet ne le possède pas. La prédiction est alors qu'au contraire cette fois l'absence de l'attribut non pertinent (plutôt que sa présence) sera perçue comme une raison pour choisir un objet. Ces deux résultats furent obtenus, confirmant ainsi les hypothèses des auteurs sur le fait qu'un même attribut peut-être utilisé comme une raison positive ou négative selon la décision que les participants ont à justifier. Cela montre bien non seulement que les participants sont prêts à utiliser des attributs qu'eux-mêmes évaluent comme étant non pertinents afin de pouvoir prendre une décision justifiable, mais également que cette utilisation est suffisamment flexible pour pouvoir servir dans les deux sens selon ce qu'ils ont besoin de justifier dans un contexte donné.

D'autres séries d'expériences ont étudié l'impact du raisonnement sur la prise en compte de facteurs non pertinents. Dans ce cas le raisonnement va modifier la pondération normalement attribuée à différents attributs, mais cette fois certains attributs seront surpondérés pour une raison bien précise. Etant donné qu'ils ont été fournis par l'expérimentateur, il faudrait justifier le fait de les ignorer. Il est alors plus simple d'en tenir compte, même si nous ne l'aurions peut-être pas fait spontanément. Dans ce cas, des attributs n'ayant qu'une pertinence limitée, ou nulle, pour le jugement vont l'influencer, et ainsi réduire la contribution d'autres facteurs plus pertinents – d'où le nom d'effet de dilution. Deux expériences ont montré que cet effet pouvait bien être attribuable au besoin de justifier le fait de ne pas prendre en compte ces attributs. Dans la première, certains participants devaient rendre des comptes (Tetlock & Boettger, 1989). Dans ce cas, ils seraient bien forcés de justifier le fait d'ignorer certains attributs. Conformément aux attentes, cette manipulation eut pour effet d'augmenter l'effet de dilution. Dans la seconde expérience, des manipulations rendaient particulièrement aisé, ou au contraire particulièrement difficile, cette même justification (Tetlock, Lerner, & Boettger, 1996). Ces

manipulations eurent l'effet escompté : elles diminuèrent l'effet de dilution dans le premier cas, et l'augmentèrent dans le second[65].

8.2.5 Des raisons spéciales

Dans une nouvelle série d'études, Simonson et Nowlis (2000) ont étudié l'impact d'un facteur modérateur sur le type de raisons employé : le besoin d'unicité (C. R. Snyder & Fromkin, 1977). Le besoin d'unicité est une variable utilisée en psychologie sociale pour mesure la tendance qu'ont les gens à se définir différemment des autres, à montrer qu'ils ne sont pas comme tout le monde. Selon les auteurs, cette variable devrait avoir une influence sur le type de raisons préféré : les personnes ayant un fort besoin d'unicité devraient préférer des raisons originales, afin de se démarquer des autres. De plus, baser son choix sur des raisons originales peut également être perçu comme une façon de montrer une certaine supériorité intellectuelle, et on peut penser que de telles raisons seront plus persuasives, dans la mesure où il s'agit d'arguments nouveaux. Afin de tester ces hypothèses ils ont comparé, dans plusieurs tâches différentes, les réponses de participants ayant un fort ou un faible besoin d'unicité (définis comme étant ceux qui se trouvent au dessus ou en dessous de la valeur médiane). Et afin de s'assurer que des différences éventuelles étaient bien dues au recours à des raisons différentes, ou pour faire naître de telles différences, ils ont également comparé une condition contrôle et une condition dans laquelle les participants devaient donner des raisons pour expliquer leur choix.

On peut souligner que certains des effets qui vont être utilisés dans cette étude comptent parmi les plus classiques de la prise de décision (aversion aux pertes et cadrage par exemple). Le plus souvent, les erreurs dans les situations de prise de décision sont expliquées par les insuffisances des mécanismes intuitifs (voir par exemple Kahneman & Frederick, 2002, 2005). En montrant que le raisonnement peut être responsable de certains des biais les plus courants, ces expériences peuvent

[65] Voir Gordon, Rozelle et Baxter (1988), Schlosser et Shavitt (2002), Siegel-Jacobs et Yates (1996), Tetlock et Boettger, (1989), Tetlock et al. (1996) et Tordesillas et Chaiken (1999) pour des résultats similaires concernant l'effet de donner des raisons sur la prise en compte de caractéristiques non pertinentes.

remettre en cause la vision classique du raisonnement dans laquelle il ne fait que corriger les biais, et non les causer.

La première expérience concerne l'effet des soldes sur les intentions d'achat. Les participants devaient tout d'abord faire un choix parmi des triplets d'items de qualité (et de prix) différents. Ensuite, ils devaient faire des choix parmi ces mêmes triplets, mais cette fois tous les produits étaient soldés de 30%. On observe alors généralement un décalage des choix vers les produits plus haut de gamme : le fait qu'ils soient soldés fournit une raison pour les acheter. Il s'agit cependant d'une raison conventionnelle. Les auteurs prédisent donc que les participants ayant un fort besoin d'unicité et devant fournir des raisons pour leur choix auront *moins* tendance à modifier leur choix en réponse aux soldes. C'est bien ce qu'ils observèrent : ils ne furent que 20% à changer d'item à cause des soldes, contre près de 40% dans les autres cas (i.e. faible besoin d'unicité et/ou pas de justification à donner).

Dans la seconde expérience, l'effet d'une autre caractéristique non pertinente[66] fut étudié. Il s'agit cette fois de la 'vantardise' (puffery), c'est-à-dire certaines prétentions des annonceurs vis-à-vis d'une caractéristique de leur produit qui sont trop grandiloquentes par rapport à l'avantage réel offert par cette caractéristique. La prédiction est cette fois que les participants ayant un fort besoin d'unicité et devant donner des raisons devraient être moins sensibles à l'effet de ces vantardises. En effet, ils devraient reconnaître qu'il s'agit là potentiellement d'une raison qui sera utilisée conventionnellement pour justifier l'achat d'un produit, et s'en servir au contraire pour justifier son rejet – dans la mesure où il n'y a pas d'autres raisons de le différencier de son concurrent. C'est bien ce qui fut observé : les participants ayant un fort besoin d'unicité et devant justifier leurs réponses furent beaucoup moins sensibles à la vantardise que ceux ayant un faible besoin d'unicité. De plus, ces derniers furent au contraire encore plus influencés par la vantardise lorsqu'ils devaient donner des raisons, confirmant l'idée que la vantardise peut être utilisée comme raison conventionnelle pour justifier un choix.

La troisième expérience concerne l'effet de compromis, et elle avait brièvement été mentionnée plus haut à cette occasion. Dans un contexte dans lequel les participants ne s'attendent pas à être évalués, mais doivent tout de même fournir

[66] Les soldes pouvaient être considérées comme non pertinentes car elles affectent également tous les produits.

des raisons, on peut s'attendre à ce que l'effet de compromis diminue. Ceci devrait être vrai en particulier des personnes ayant un fort besoin d'unicité, pour qui des explications en termes de 'cette option est moyenne' devraient être particulièrement peu attirantes. Les résultats ne confirmèrent la prédiction que partiellement. Conformément à ce qui avait été observé par Briley et al. (2000) chez leurs participants Nord-Américains, le fait de devoir donner des raisons diminua l'effet de compromis, mais l'effet ne fut pas modéré par le besoin d'unicité. Par contre les participants ayant un fort besoin d'unicité montrèrent moins d'effet de compromis (et montrèrent même un effet négatif dans la condition 'raison') dans tous les cas.

Les deux expériences suivantes concernent l'effet d'aversion aux pertes ('loss aversion', voir (Kahneman & Tversky, 1979, 1991). L'effet peut s'illustrer par la comparaison entre les problèmes suivants (qui sont ceux utilisés dans l'expérience, voir Simonson & Nowlis, 2000, p.57) :

> Problem 1.—Option A: $30 for sure; Option B: 50 percent to lose $100 and 50 percent to win $300.
> Problem 2.—Option A: $230 for sure; Option B: 50 percent to win $100 and 50 percent to win $500.

Alors que ces deux comparaisons sont identiques à un facteur fixe près (200$ de plus pour chaque somme dans le second problème), l'aversion aux pertes prédit que les participants choisiront plus souvent l'option A dans le problème 1 (qui représente un gain alors que l'autre option comprend un risque de perte) que dans le problème 2 (dans lequel les deux options ne peuvent qu'entraîner des gains). Etant donné qu'il s'agit là d'une solution courante, dont on peut penser qu'elle s'assortit le plus souvent d'une raison conventionnelle (éviter une perte potentielle), les personnes ayant un fort besoin d'unicité et devant donner des raisons pour leurs réponses devraient être moins sensibles à l'aversion aux pertes.

La première expérience compara les deux paires de problèmes mentionnées ci-dessus. Les résultats confortèrent les hypothèses. Dans tous les autres cas, les participants montrèrent une aversion aux pertes (en choisissant plus souvent l'option A dans le problème 1 que dans le problème 2). Mais les participants ayant un fort besoin d'unicité et devant donner des raisons eurent le comportement *inverse*. Etant donné que cela viole un des résultats les mieux établis de l'économie

comportementale (l'aversion aux pertes), les auteurs ont cherché à répliquer ce résultat dans une seconde expérience. Un autre objectif était d'éliminer une explication alternative possible selon laquelle la raison de choisir l'option A dans le problème 1 était beaucoup plus faible que celle de choisir la même option dans le second problème (30$ étant très inférieur à 230$). Cela pourrait en effet expliquer les résultats sans qu'ils ne concernent des différences de raisons portant sur l'aversion aux pertes (portant sur les différences dans l'option B). A cette fin, un nouveau problème fut introduit qui égalisait l'option A (à 230$) tout en maintenant une option B avec un risque de perte. Les résultats de l'expérience 1 furent répliqués, y compris l'inversion du phénomène d'aversion aux pertes chez les participants ayant un fort besoin d'unicité et devant justifier leurs décisions (avec cette fois 22% de différence en faveur de l'option impliquant une perte potentielle). Ces résultats furent confirmés pour le nouveau problème – à nouveau, les participants ayant un fort besoin d'unicité et devant justifier leurs réponses furent les seuls à montrer un comportement opposé à l'aversion aux pertes. On peut maintenant se demander quelles sont les limites de l'explication du phénomène d'aversion aux pertes par le choix basé sur des raisons. Les auteurs eux-mêmes indiquent que l'explication pourrait ne s'appliquer qu'aux cas dans lesquels les participants doivent faire des choix (voir plus bas pour une explication des différences entre les situations de choix et d'évaluation). Quoiqu'il en soit il s'agit là d'un des phénomènes centraux de l'économie comportementale, et le fait qu'il puisse recevoir une explication dans ce cadre est réellement intriguant.

La dernière expérience de cet article se penche sur un autre phénomène basé sur l'aversion aux pertes : le cadrage ('framing'). Il s'agit d'une comparaison entre problèmes structurellement identiques (les probabilités décrites y sont les mêmes), mais qui mettent l'accent sur des gains ou des pertes potentielles (Tversky & Kahneman, 1981). Lorsque les problèmes sont présentés dans le cadre de gains potentiels, les participants tendent à préférer la solution sûre, alors que lorsque le cadrage est en termes de pertes, ils tendent à préférer l'option plus risquée. Le problème le plus connu (dont ceux utilisés par Simonson et Nowlis sont des analogues) est celui de la maladie tropicale (Tversky et Kahneman, 1981, p.453) :

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to

combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

	Gain frame
If Program A is adopted, 200 people will be saved.	72%
If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.	28%
Which of the two programs would you favor?	

	Loss frame
If Program C is adopted 400 people will die.	22%
If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.	78%
Which of the two programs would you favor?	

A nouveau, on peut expliquer ces différences en termes de raisons. Lorsque le problème est cadré en termes de gains, les raisons les plus facilement disponibles, ou les plus convaincantes, seraient du type ‘au moins, on aura sauvé 200 personnes’, alors que cadré en termes de pertes, elles seraient du type ‘on ne peut pas laisser 400 personnes mourir’. L’hypothèse, à nouveau, est que les participants ayant un fort besoin d’unicité et devant justifier leurs réponses préféreront donner des raisons différentes de ces raisons plus conventionnelles. Si ces raisons ont autant de chances d’aller dans un sens que dans l’autre, cela devrait entraîner une réduction des effets de cadrage. C’est bien ce qui fut observé, avec une interaction entre les trois facteurs (cadre, besoin d’unicité et justification) significative indiquant que les participants ayant un fort besoin d’unicité et donnant des raisons étaient moins sensibles aux effets de cadrage.

8.2.6 Effets de cadrage

On peut mentionner ici plusieurs autres résultats qui soutiennent, plus ou moins directement, une interprétation des effets de cadrage en termes de choix basé sur des raisons. Le premier est une réplication partielle des effets de devoir donner des justifications observé par Simonson et Nowlis (2000), le second est une étude sur

l'effet de cadrage lors de la résolution de problème en groupe, et la troisième une expérience dans laquelle des raisons autres que celles rendues pertinentes par les cadres sont introduites.

LeBoeuf et Shafir (2003) ont conduit une expérience (expérience 1) utilisant plusieurs problèmes de cadrage. Leur objectif était de tester l'idée que les participants réfléchissant davantage devraient être moins sensibles à l'influence du cadrage. Afin d'étudier cela, ils ont d'une part mesuré le besoin de cognition ('need for cognition', Cacioppo & Petty, 1982), une variable représentant le goût pour la réflexion et la propension à s'y engager, et d'autre part demandé aux participants d'une condition de justifier leurs réponses. Si les effets de cadrage résultent bien d'un choix basé sur des raisons, ces manipulations ne devraient pas être particulièrement efficaces. S'il est possible qu'en réfléchissant davantage les participants en viennent à trouver des raisons différentes de celles qu'ils envisagent le plus naturellement (et qui mènent aux effets de cadrage), il est également possible que des participants pour lesquels ces raisons n'étaient pas les plus pertinentes immédiatement y pensent et les trouvent convaincantes. De plus, le temps supplémentaire peut n'être passé qu'à ressasser les raisons trouvées initialement, n'entraînant aucun changement si ce n'est peut-être une augmentation de la confiance.

Conformément à ces prédictions (et non aux leurs propres), LeBoeuf et Shafir n'observèrent aucun effet global ni du besoin de se justifier, ni du besoin de cognition sur la sensibilité au cadrage. Parmi les interactions portant sur des problèmes particuliers, le fait de devoir donner des justifications augmenta l'effet de cadrage dans deux cas et le diminua dans un. Ils passent également en revue la littérature sur les effets sur le cadrage de l'obligation de se justifier, et en concluent que les effets généralement trouvés sont faibles (non significatifs, ou persistance de l'effet de cadrage dans la condition justification) ou suspects (non réplication de l'effet de cadrage dans la condition contrôle). Etant donné l'effet important du besoin d'unicité démontré par Simonson et Nowlis (2000) lorsque les participants doivent justifier leurs réponses, il n'est guère surprenant que les résultats qui ne prennent pas cette variable en compte ne soient pas concluants (voir cependant Igou & Bless, 2007, expérience 2). Par exemple, on devrait s'attendre à des variations non seulement en fonction de la répartition du besoin d'unicité dans la population, mais

également de l'aspect plus ou moins conventionnel des raisons créant l'effet de cadrage.

Plusieurs expériences ont étudié les effets de cadrage lorsque les participants font face aux problèmes en groupe. Dans la mesure où les choix individuels sont basés sur des raisons, il est intéressant de mesurer les effets que peuvent avoir ces raisons lorsque les participants en débattent. On peut mentionner ici deux études qui se sont penchées sur ce problème, celle de McGuire et al. (1987) et celle de Paese et al. (1993).

Paese et al. (1993) ont comparé les effets de cadrage chez les individus et les groupes en donnant à des individus une version cadrée en termes de gains, avant de leur redonner les mêmes problèmes, mais cette fois en groupe, et ils ont fait de même avec des problèmes identiques mais cadrés en termes de pertes. En plus du problème classique de la maladie tropicale, évoqué plus haut, trois autres problèmes furent utilisés. Les résultats étaient très hétérogènes. Dans un cas, la polarisation prédite fut nettement observée : les groupes résolvant des problèmes en termes de gains montrèrent plus d'aversion au risque que les participants isolés, alors que ceux confrontés à des problèmes en termes de pertes montrèrent au contraire plus de goût pour le risque. Par contre, dans les autres cas, ce qui fut observé peut-être qualifié d'avantage rhétorique pour une raison, et ce quelque soit le cadrage. Par exemple, le troisième problème montra un décalage vers la sécurité en groupe, et ce quelque soit le cadre, alors que le quatrième problème montrait lui un décalage dans la direction de la prise de risque, à nouveau quelque soit le cadre. Les auteurs expliquent cette différence par le contenu des problèmes : alors que le troisième problème impliquait un problème de santé (tenter ou non une opération chirurgicale risquée), le quatrième impliquait un problème monétaire (un pari).

On peut alors conceptualiser les résultats de la façon suivante. Les effets de cadre observés en individuel résultent du fait que le cadre rend plus ou moins pertinentes des raisons différentes. Il peut également les rendre plus convaincantes, mais cela n'est pas évident, et pourra être contrebalancé par d'autres facteurs – comme le contenu du problème. Etant donné que les participants ne se trouvent pas, en individuel, dans un contexte propice à une utilisation très poussée du raisonnement, il est fort possible qu'ils se contentent des raisons qui se présentent à

eux en premier, sans chercher à les tester plus avant (voir le principe de satisficing appliqué au raisonnement dans la section 3.1). Ces effets de pertinence n'ont pas toujours le même résultat : ils n'empêchent pas certains participants d'utiliser d'autres raisons, ce qui est indiqué par le fait que les réponses ne sont jamais au plafond. Donc, lorsque les participants se retrouvent en groupe, plusieurs raisons seront en compétition. Alors la simple pertinence ne jouera qu'un rôle mineur, c'est plus la force de conviction d'une raison par rapport aux autres qui sera importante. Et comme la force de conviction ne dépend pas uniquement du cadre, mais du contenu du problème (et d'autres facteurs tels que la culture des participants, le contexte plus général dans lequel ils se trouvent, etc.), les résultats en groupe peuvent aller dans une direction différente de celle causée par les effets de cadrage. Les auteurs eux-mêmes suggèrent une explication de leurs résultats dans les termes de la théorie des arguments persuasifs (Vinokur, 1971; Vinokur & Burnstein, 1978). Selon cette théorie, un des éléments déterminants de la prise de décision en groupe est la qualité des arguments présentés par les différents participants. Si cette explication est tout à fait compatible avec la théorie argumentative, elle serait beaucoup plus dure à réconcilier avec des théories plus normatives de la prise de décision. En effet, certains facteurs qui influencent la pertinence et la force de persuasion des arguments ne devraient pas jouer de rôle selon les théories normatives (tel que le cadre, ou certains aspects du contenu des problèmes).

L'explication des effets de cadrage offerte par la théorie des arguments persuasifs est renforcée par l'expérience de McGuire et al. (1987). Cette expérience est similaire à la précédente : les participants devaient d'abord résoudre des problèmes de prise de décision (soumis à des effets de cadrage) individuellement, puis en groupe. Une autre variable était cependant étudiée : l'effet du mode de communication. Alors que dans la condition 'conversation' les membres du groupe étaient assis autour d'une table et pouvaient débattre normalement, ceux de la condition 'ordinateur' devaient passer par un logiciel leur permettant de taper leurs messages, messages qui étaient montrés en directs sur les ordinateurs des autres participants. En plus de ces différences, les résultats révélèrent un autre facteur ayant pu jouer un rôle important : la manipulation visant à créer un effet de cadrage ne fut pas efficace en individuel. Le débriefing indiqua qu'une part considérable des participants avait simplement conceptualisé la tâche (qui était une tâche monétaire,

au contenu donc très différent de celui de la maladie asiatique par exemple) comme une simple tâche de mathématiques, de calcul. Ce résultat est intéressant car malgré lui, les effets de cadrage s'avérèrent être très forts dans la condition 'discussion'. Une analyse des réponses des participants avant et après la discussion indiqua que le schème décrivant les données était « une norme gagne » (Davis, 1973). Dans ce schème, une réponse est soutenue par une norme (une raison) qui est tellement convaincante qu'il suffit qu'un seul membre du groupe la présente pour que le groupe entier l'adopte. Dans ce cas, les normes 'gagnantes' étaient celles prédites par la théorie des perspectives ('prospect theory'), c'est-à-dire une attirance pour la sécurité dans le cas des gains, et pour le risque dans le cas des pertes.

La comparaison avec la condition 'ordinateur' renforça l'interprétation en termes d'arguments persuasifs. Dans cette condition aucun effet de cadrage n'émergea, au contraire donc de la condition 'discussion' qui vit naître de forts effets de cadrage. Pour expliquer cette différence, les auteurs ont analysé le contenu des conversations dans les deux conditions. Dans les deux cas, le nombre d'énoncés contenant l'avis des membres du groupe était similaire, ce qui signifie que le simple fait de savoir ce que pensent les autres n'avait pas d'influence réelle. Par contre, dans la condition 'discussion', les auteurs observèrent significativement plus d'arguments. Ils expliquent donc la différence entre les conditions, et l'émergence des effets de cadrage dans la condition 'discussion', par ces échanges d'arguments.

Finalement, un troisième argument qu'on peut invoquer pour soutenir l'interprétation des effets de cadrage en termes de la théorie des choix basés sur des raisons est que l'introduction de raisons différentes de celles rendues pertinentes par le cadre influence les réponses. Si les différences entre cadres étaient dues à des évaluations différentes faites des gains et des pertes (selon la théorie des perspectives, de laquelle découle l'aversion aux pertes), alors des modifications minimes des quantités présentées ne devraient avoir aucun effet sur les résultats. Fulginiti et Reyna (rapporté dans Reyna & Brainerd, 1995) ont cependant observé un résultat qui contredit cette prédiction. Ils avaient pour cela repris le problème de la maladie tropicale, en en modifiant légèrement les paramètres : dans certains problèmes, les paramètres étaient modifiés dans une direction cohérente avec le cadrage (sauver 201 personnes / avoir une chance sur trois que 597 personnes meurent), dans d'autres ils l'étaient dans la direction opposée (sauver 199 personnes /

avoir une chance sur trois que 603 personne meurent). Ces différences minimes dans les quantités présentées ne devraient avoir aucun effet décelable si les participants utilisaient les mécanismes décrits dans la théorie des perspectives. Par contre, ces faibles différences rendent de nouvelles raisons saillantes : 201 est plus qu'un tiers de 600 (et 200 plus qu'un tiers de 597), ce qui donne une raison supplémentaire de prendre la décision déjà soutenue par le cadrage, et vice versa pour la modification opposée. Les différences entre ces deux conditions et une condition contrôles furent significatives : l'ajout de raisons allant dans le sens du cadrage renforça les effets de cadrage et de raisons allant dans le sens opposé les diminua.

8.2.7 Inversion de préférence

L'évaluation des préférences est un outil essentiel des économistes : sans ces évaluations, impossible de prédire le comportement. Pour que ces mesures soient utiles, elles doivent être relativement robustes face à des variations non pertinentes pour les préférences elles-mêmes, telles que des variations dans la façon de les évaluer. Or les psychologues ont découvert de très nombreuses violations de ce principe (appelé le principe d'invariance de procédure) : les gens évaluent différemment leurs préférences selon le type d'échelle utilisé, ou la question qui est posée, à tel point qu'on peut observer dans certains cas des *inversions de préférence* (voir Hsee, Loewenstein, Blount, & Bazerman, 1999; Tversky, Sattath, & Slovic, 1988 pour des revues). Il s'agit de situations dans lesquelles les préférences s'inversent selon le moyen par lequel elles sont mesurées. Il ne s'agit pas d'un phénomène causé par un seul mécanisme. Il est cependant important de le distinguer d'une autre forme d'inversion des préférences causée elle non pas par les moyens de mesure, mais par le délai séparant la survenue de différents événements (voir par exemple Kirby & Herrnstein, 1995). Dans ce cas, le mécanisme causal principal est connu : il s'agit de 'l'hyperbolic discounting' (actualisation hyperbolique), un mécanisme partagé par tous les animaux chez qui ce type de phénomène a été étudié, et qui est distinct du raisonnement. Nous ne nous intéresserons pas à cet effet car l'objectif de cette section est de montrer que c'est le raisonnement qui cause une partie considérable des phénomènes d'inversion de préférence.

Un des premiers phénomènes liés à l'inversion de préférence à être expliqué dans le cadre du choix basé sur des raisons a été démontré par Slovic (1975). Il s'agit d'une tendance à choisir, parmi deux options équivalentes, celle qui l'emporte sur l'attribut le plus important. Afin de s'assurer que les options étaient bien équivalentes, Slovic avait commencé par demander aux participants de compléter des problèmes comme le suivant :

	Gift package A	Gift package B
Cash	_	$20
Coupon book worth	$32	$18

Ils devaient donner la valeur qui rendrait les deux ensembles équivalents. Une semaine plus tard, on leur demandait de choisir parmi les deux ensembles, après que la valeur qu'ils avaient eux-mêmes évaluée ait été insérée. On leur demandait également de juger la caractéristique qui était la plus importante à leurs yeux. Plutôt que de ne pas être différentes du hasard, les réponses favorisaient très nettement l'alternative qui était la plus forte sur l'attribut jugé le plus important (88%). Tversky et al. (1988) ont ensuite confirmé ce résultat par le biais de résultats interindividuels. Il est possible d'expliquer ces résultats en termes de facilité de justification : les justifications dépendant de l'attribut jugé le plus important sont plus pertinentes, ou sont jugées comme étant plus convaincantes, que les autres, et elles font donc pencher la balance en direction du choix étant supérieur pour cet attribut. Comme le disent Tversky et al. : « it provides a compelling argument for choice that can be used to justify the decision to oneself as well as to others » (p.372).

Parmi les très nombreuses inversions de préférences observées dans la littérature, un bon nombre peut s'expliquer facilement par l'influence différentielle de différentes raisons. Une des théories principale expliquant ces phénomènes se base sur 'l'*évaluabilité*' (Hsee, 1996b; Hsee et al., 1999). Bien qu'elle puisse s'appliquer à d'autres types d'inversion de préférences, elle est concernée avant tout par les inversions entre des situations d'*évaluation jointe* et d'*évaluation séparée*. Voici un exemple (il s'agit de l'expérience 1 de Hsee, 1996b). Les participants devaient évaluer le prix qu'ils seraient prêts à payer pour des articles. Dans la condition d'évaluation jointe, deux articles étaient présentés ensemble. Dans les

conditions d'évaluation séparée, ces articles étaient présentés à des participants différents qui ne pouvaient donc pas effectuer de comparaison entre les deux. Voici une telle paire d'articles (il s'agit de dictionnaires musicaux) :

	Dictionary A	Dictionary B
Year of publication:	1993	1993
Number of entries:	10,000	20,000
Any defects?	No, it's like new.	Yes, the cover is torn; otherwise it's like new

Les résultats (les sommes que les participants seraient prêts à payer) indiquèrent une préférence pour le dictionnaire A en condition d'évaluation séparée mais une préférence pour B en condition d'évaluation jointe (différence significative). L'auteur explique cette différence par le fait que l'attribut 'nombre d'entrées' est difficilement évaluable lorsqu'on n'a pas de point de comparaison (qui sait combien d'entrées est supposé contenir un dictionnaire musical ?), et qu'il joue donc un rôle mineur en évaluation séparée. Cependant, il devient possible de l'évaluer lorsqu'un point de comparaison apparaît, et il fait alors pencher la balance dans l'autre direction en évaluation jointe. Au contraire, l'attribut 'défauts' serait facile à évaluer même sans point de comparaison direct.

Il apparaît dans cet exemple que le concept 'd'évaluabilité' pourrait tout aussi bien être de la 'justifiabilité' : un attribut joue un poids plus ou moins important car on peut justifier l'influence de cet attribut dans la décision. Cette interprétation en termes de justifications offre plusieurs avantages. Elle permet d'incorporer les résultats dans le cadre du choix basé sur des raisons, et ainsi de faire de nouvelles prédictions. Par exemple, l'option choisie devrait être perçue comme étant moins facilement critiquable, plus facile à justifier. On pourrait également s'attendre à observer des effets du fait de devoir donner des raisons – peut-être modérés par le besoin d'unicité des participants[67]. Et ce cadre offre également une bonne explication pour l'influence 'indue' de l'attribut qui est le moins pertinent dans les conditions

[67] Les deux expériences ayant manipulé le fait de devoir rendre des compte dans des situations d'inversion de préférence ne concernaient pas cette différence entre modes d'évaluation séparé et joint (Selart, 1996; Simonson & Nye, 1992).

d'évaluation séparée. Il devient en effet possible de recourir alors à l'effet de dilution. Il y a deux raisons pour lesquelles les informations fournies par l'expérimentateur sont susceptibles d'avoir un poids trop important dans le processus de prise de décision. D'une part le fait qu'elles aient été communiquées les rend plus pertinentes. D'autre part les rejeter nécessite une bonne raison, ce qui peut demander des efforts tels qu'il est plus simple de la prendre en compte (voir section 8.2.4). Dans l'exemple ci-dessus, il s'agit de l'information qu'un des dictionnaires a une couverture déchirée. Bien que cette information ne soit pas aussi peu pertinente que les informations typiques de l'effet de dilution, son mode de présentation peut lui accorder une importance démesurée par rapport à sa pertinence réelle[68].

Ces explications basées sur la facilité avec laquelle on peut justifier l'influence de différents facteurs s'appliquent aussi bien aux autres résultats de Hsee, (1996b), et également aux autres expériences expliquées par l'hypothèse d'évaluabilité. Par exemple, Hsee (1998, expérience 2) a comparé les situations suivantes :

> Imagine the following scenario: It is summer in Chicago. You are on the beach at Lake Michigan. You find yourself in the mood for some ice cream. There happens to be an ice cream vendor on the beach. She sells Haagen Dazs ice cream by the cup. For each serving, she uses [a 10 oz cup and puts 8 oz of ice cream] {a 5 oz cup and puts 7 oz of ice cream} in it.

Les produits décrits sont donc une coupe de glace de 10 onces n'en contenant que 8 (sous-remplie) et une de 5 en contenant 7 (sur-remplie). Dans la situation d'évaluation séparée, les participants partent probablement d'une estimation similaire (le prix d'une coupe de glace, qui ne varie pas réellement entre 7 et 8 onces), puis l'ajustent vers le haut ou vers le bas selon que la coupe est sous- ou sur-remplie, puisqu'il s'agit là de la seule autre information qui leur est fournie. Par contre, dans la situation d'évaluation jointe, ils peuvent comparer directement les quantités de glace. Dans ce cas, leurs intuitions concernant l'utilisation du remplissage comme raison font face à leur compréhension qu'ils passeraient pour des incompétents s'ils

[68] Un moyen de tester cela serait de présenter des items réels plutôt que des descriptions.

attribuaient un prix supérieur à une coupe contenant moins de glace. Nous reviendrons plus bas sur le fait qu'un élément qui indique que les participants sont plus intéressés à justifier leurs décisions qu'à prendre de bonnes décisions est le fait que les décisions qu'ils prennent ainsi ne maximisent souvent pas leur satisfaction. En particulier les décisions prises, comme c'est le cas dans l'évaluation des glaces en mode joint, principalement afin de s'assurer qu'ils n'aient pas l'air incompétent ont des chances de ne pas maximiser leur satisfaction.

Un autre exemple d'inversion qui pourrait avoir des conséquences similaires est le suivant (expérience de Bazerman, Loewenstein, & White, 1992). Une décision de justice est rendue à propos d'un litige qui vous opposait à votre voisin. Voici les deux décisions possibles, les participants devant juger celle qui les contenterait le mieux :

Décision 1 :	\$600 pour soi et \$800 pour le voisin
Décision 2 :	\$500 pour soi et \$500 pour le voisin

Dans ce cas, l'option 2 est préférée en évaluation séparée, mais la première l'est en évaluation jointe. Dans le cas de l'évaluation séparée, l'intuition qu'une répartition égale est bonne fait pencher la balance. Par contre, dans le cas de l'évaluation jointe, on paraîtrait incompétent à préférer gagner moins d'argent. Il est cependant loin d'être évident que, quoiqu'ils en disent au moment de répondre à la question, les participants soient réellement plus heureux avec la décision 1 qu'avec la 2.

On retrouve encore des effets similaires dans des expériences de Kahneman et Ritov (1994) et d'Irwin, Slovic, Lichtenstein et McClelland (1993) rapportées dans Hsee et al. (1999). Voici des exemples de problèmes utilisés, les participants devant dire combien ils donneraient pour lutter pour chaque cause. A chaque fois, la première option présentée est celle pour laquelle les gens donneraient plus en mode d'évaluation jointe, et la seconde en mode d'évaluation séparée :

1) Skin cancer from sun exposure common among farm workers.
2) Several Australian mammal species nearly wiped out by hunters.

1) Multiple myeloma among the elderly.
2) Cyanide fishing in coral reefs around Asia.

1) Improving the air quality in Denver.
2) Adding a VCR to your TV.

On voit bien qu'à chaque fois les participants risqueraient, non pas tant de paraître incompétent, mais ne respectant pas certaines normes morales. Les normes morales (la vie humaine vaut plus que la vie d'animaux, le bien commun est supérieur à mes désirs personnels) qui font pencher la balance dans les évaluations jointes ne sont généralement pas spontanément activées par l'évaluation des options séparées. A nouveau, il est loin d'être clair que les participants seraient plus satisfaits s'ils respectaient les choix faits en évaluation jointe (qui est pourtant la norme des théories de la décision, qui poussent toujours les gens à comparer le plus d'options possibles). Par exemple, il est très probable que l'achat d'un magnétoscope apporte plus de satisfaction que la modeste contribution qui pourrait être faite avec la même somme à l'amélioration de la qualité de l'air[69].

Dans une série d'expériences, Hsee et Zhang (2004) ont démontré que ces phénomènes pouvaient mener à des choix sous-optimaux au sens ou ils ne maximisent pas la satisfaction. Leur hypothèse est que les attributs que l'on peut distinguer quantitativement joueront un rôle plus (trop) important dans les situations d'évaluation jointe, alors que les attributs qualitatifs y joueront un rôle moins important, et que cela amènera les participants à choisir des options qui ne sont en fait pas celles qui leur apporteraient la plus grande satisfaction. Dans leur première expérience, les participants devaient se mettre à la place d'une personne ayant écrit un recueil de poésie qu'elle essaie de vendre sur le campus. On leur demandait d'indiquer à quel point différents scénarios les rendraient plus ou moins heureux. Dans un scénario, ils n'avaient vendu aucun livre, dans un autre 80, et 160 et 240 dans les deux derniers. Lorsque ces différentes options étaient évaluées séparément, il y avait une grosse différence entre aucun et les trois quantités positives, qui

[69] Cela ne veut cependant pas dire que ces décisions ne sont pas meilleures pour la société, au contraire, c'est là justement tout le point des normes morales.

n'étaient pas significativement différentes entre elles. Par contre, lorsque les participants évaluaient deux options simultanément, la différence entre aucun et 80 fut largement sous-estimée, alors que les autres différences (entre 80 et 160, et 160 et 240) furent largement surestimées. Ceci conforte bien la prédiction que la différence qualitative (entre aucun et 80) joue un rôle qui n'est pas assez important en évaluation jointe alors qu'au contraire les différences quantitatives y jouent un rôle trop important.

La seconde expérience est proche de la première, mais elle inclut également une condition dans laquelle l'évaluation par les participants d'une situation réellement vécue était également mesurée. Ceci permet de s'assurer que les estimations séparées prédisent mieux les sentiments réels que les estimations jointes. La tâche consistait à lire une quantité variable de mots de valence variable (10 ou 25 mots à valence positive ou négative). Dans la condition 'expérience', les participants lisaient ces mots puis indiquaient comment ils se sentaient. Dans la condition 'séparée', les participants lisaient une description d'une seule tâche (par exemple lire 10 mots de valence positive) et indiquaient comment ils se sentiraient. Dans la condition 'jointe', les participants devaient prédire leur satisfaction pour chacune des quatre tâches. Les résultats montrèrent une parfaite corrélation entre évaluation après expérience et évaluation en présentation séparée. Par contre, et conformément aux prédictions, les participants de la condition 'jointe' surévaluèrent l'influence des différences quantitatives entre 10 et 25 mots (influence qui était en fait non significative dans le cas de l'expérience vécue).

La troisième expérience implique la comparaison de deux options qui diffèrent par deux attributs. Il s'agit d'un choix entre une tâche pénible (décrire par écrit un échec personnel) accompagnée d'une récompense plus importante (un chocolat de 15 grammes) et une tâche agréable (décrire par écrit un succès personnel) accompagné d'une récompense moins importante (un chocolat de cinq grammes). Dans une condition ('choix pour soi'), les participants avaient le choix entre les deux tâches. Avant de choisir, ils devaient évaluer leurs deux composants séparément – prédisant la satisfaction qu'ils apporteraient. Puis, après avoir accompli la tâche (et mangé le chocolat), ils indiquaient leur satisfaction. Dans une autre condition ('choix pour autre'), les participants devaient choisir pour une autre personne, afin de maximiser sa satisfaction. Enfin, dans les deux dernières conditions, les participants ne devaient faire qu'une des deux tâches, puis évaluer leur satisfaction. Les résultats

confirmèrent les prédictions. Les participants devant choisir attribuèrent trop d'importance à la taille du chocolat (qui n'avait en fait pas d'influence significative sur la satisfaction), ce qui les conduisit à faire des choix sous-optimaux : la majorité d'entre eux (plus de 60%, qu'il s'agisse d'un choix pour eux ou pour un autre) choisit la combinaison d'une tâche désagréable et d'une récompense plus importante, alors qu'ils auraient été plus satisfaits en choisissant l'autre option[70]. En effet, la satisfaction des participants ayant choisi l'autre option (tâche agréable et plus petite récompense) était beaucoup plus importante, de même que celle des participants n'ayant pas eu le choix.

Ces exemples montrent bien à quel point l'utilisation d'attributs non pertinents pour prendre des décisions peut amener à faire des choix sous-optimaux. On peut alors se demander pourquoi est-ce que les gens surpondèrent ainsi les informations quantitatives. Une explication est qu'elles fournissent de meilleures raisons, qu'elles sont plus faciles à exprimer (un chocolat est trois fois plus gros que l'autre, ça doit être important, je paraîtrais stupide si je prenais le petit chocolat simplement parce qu'il me faut accomplir une tâche un peu déplaisante). Ceci amène ces attributs à jouer un rôle trop important dans le choix, de la même manière que, par exemple, les éléments plus facilement verbalisables peuvent être surpondérés dans des circonstances similaires (voir Schooler & Engster-Schooler, 1990; Sengupta & Fitzsimons, 2004). Il serait difficile d'expliquer ces effets dans les termes d'une théorie plus normative du raisonnement. On pourrait arguer que la théorie argumentative ne fournit pas une explication détaillée de pourquoi les informations quantitatives font de meilleures raisons. Cela n'empêche qu'elle prédit que les meilleures raisons devraient jouer un rôle dans la prise de décision, en tant que justifications. Or ce rôle serait lui aussi à expliquer dans le cadre d'une théorie normative du raisonnement. Cette dernière devant recourir à une explication ad hoc dans ce cas, on est en droit de préférer l'explication dans les termes de la théorie argumentative, ou du choix basé sur des raisons.

[70] On peut noter que les participants d'une autre condition firent la même erreur lorsqu'ils devaient choisir pour quelqu'un d'autre. Ce n'est guère étonnant dans la mesure où ces participants commençaient sûrement par ce demander ce qu'ils choisiraient eux-mêmes avant de déterminer la tâche (et la récompense) qu'ils attribuaient à une autre personne.

On peut finalement mentionner une étude sur le même sujet portant sur la différence de facilité de justification entre les alternatives 'hédonistes' et les alternatives 'utilitaires' (Okada, 2005). L'auteure part de l'observation que les achats, ou décisions plus généralement, hédoniques sont plus difficiles à justifier que les décisions utilitaires, car elles sont associées à un sens de culpabilité, ou qu'elles sont plus difficiles à quantifier (voir par exemple Loewenstein & Prelec, 1998; Thaler, 1980). Sur cette base, elle fait l'hypothèse que dans une situation d'évaluation jointe, les alternatives hédoniques auront tendance à être choisies moins souvent qu'elles n'auraient dû l'être à en croire leurs évaluations séparées. La première expérience est très naturelle puisqu'il s'agit des choix de consommateurs dans un restaurant. Ce restaurant ne servait qu'un ou deux desserts. La consommation fut donc comparée entre un soir durant lequel il ne servait qu'un dessert qualifié d'hédonique (un cheese cake d'apparence grasse, sucrée, et goûteuse), un soir ou au contraire le seul dessert était utilitaire (un cheese cake 'light'), et enfin un soir ou les deux options étaient disponibles. Bien que la différence dans la consommation de dessert entre les deux soirs pour lesquels un seul dessert était présenté n'ait pas été significative, lorsque les clients pouvaient choisir entre les deux, ils choisirent plus souvent la version légère, utilitaire.

Dans la deuxième expérience, les participants devaient évaluer et choisir entre deux options : un coupon pour 50$ de courses au supermarché (option utilitaire), ou un coupon pour 50$ dans un restaurant (option hédonique). Dans un ordre contrebalancé entre les participants, et à une semaine d'intervalle, ils devaient : évaluer la satisfaction de recevoir un coupon, faire la même évaluation pour l'autre coupon, choisir un des deux coupons, et dire quel coupon ils aimeraient qu'un ami choisissent pour eux. Cette dernière condition était ajoutée afin de supprimer le besoin de justification lié à la culpabilité de choisir l'option utilitaire. Les résultats furent parfaitement conformes aux prédictions. Alors que les participants estimaient qu'ils seraient significativement plus satisfaits par le coupon pour le restaurant que pour le supermarché, ils choisirent en majorité de prendre celui pour le supermarché. Par contre, débarrassés de la contrainte de culpabilité, plus de la moitié préféraient qu'un ami choisissent pour eux l'option hédonique.

L'hypothèse testée dans la troisième expérience est la suivante. La valeur du temps est plus ambiguë que la valeur de l'argent (voir Okada & Hoch, 2004). Il est donc plus aisé de justifier une dépense de temps qu'une dépense d'argent. Si les gens

ressentent plus le besoin de justification dans le cas des décisions hédoniques, ils devraient plus facilement dépenser du temps que de l'argent (car la dépense peut être artificiellement considérée comme moindre dans le cas du temps). Au contraire, ceux devant justifier des décisions utilitaires devraient plus facilement dépenser de l'argent (car ils peuvent justifier leur décision de toute façon). Dans cette étude, des personnes ayant acheté un appareil photo donné remplissaient un questionnaire concernant leur achat. Ils devaient indiquer s'il s'agissait plus pour eux d'un achat hédonique ou utilitaire, et le temps qu'ils seraient prêts à passer pour aller dans un autre magasin afin de trouver le même modèle avec une réduction de 50$. Les résultats montrèrent que les personnes qui considéraient davantage l'achat comme hédonique étaient prêtes à passer significativement plus de temps pour avoir la réduction, ce qui implique qu'au contraire celles qui considéraient l'achat comme plus utilitaire étaient davantage prêtes à dépenser de l'argent (en n'obtenant pas la réduction). On peut cependant donner à ces résultats une interprétation plus simple : les participants pour qui l'achat était le plus hédonique auraient eu plus de mal à justifier le fait de ne passer que peu de temps pour aller chercher la réduction, et ils tendaient donc à indiquer une durée supérieure. L'expérience quatre répliqua ce résultat dans un contexte plus contrôlé et avec différents items.

Ici encore, la facilité avec laquelle on peut justifier telle ou telle option amène à faire des choix sous-optimaux, en particulier à ne pas choisir suffisamment les alternatives hédoniques lorsqu'elles sont confrontées à des alternatives utilitaires.

8.2.8 Rationalisme naïf

Hsee a également étudié un autre biais dans la façon dont les décisions sont prises qui est très pertinent ici, ce qu'il appelle le 'rationalisme naïf' : « We propose that decision-makers have a tendency to resist affective influence, and to rely on rationalistic attributes to make their decisions. » (Hsee, Zhang, Yu, & Xi, 2003), p.16). A nouveau, ces résultats s'interprètent naturellement en termes de justifications : les gens prennent certaines décisions car ils pensent qu'elles seront plus faciles à justifier, ici au moyen d'attributs paraissant 'rationnels'.

La première catégorie d'effets est rapportée sous le nom 'd'économisme naïf', une tendance, déjà observée dans des expériences précédentes, à accorder trop

de poids à la valeur monétaire par rapport à d'autres attributs qui en fait jouent un rôle plus important pour la satisfaction (sur l'économisme naïf, voir également le phénomène lié de maximisation du moyen [Hsee, Yu, Zhang, & Zhang, 2003]). Hsee et al. commencent par décrire une expérience de Hsee (1999) dans laquelle les participants devaient soit évaluer soit choisir entre les deux options suivantes : un petit chocolat en forme de cœur (0,5 once, 0,50$) et un plus gros chocolat en forme de cafard (2 onces, 2$). Les participants furent plus nombreux à prédire (probablement à raison si on en croit les résultats de la psychologie du dégoût) qu'ils préféreraient le petit chocolat en forme de cœur qu'ils ne furent à le choisir.

La première expérience de Hsee et al. (2003) est très proche de celle-ci. Il s'agissait là d'un coupon pour quatre dîners. Dans un cas, la valeur des quatre dîners augmentait (une qualité) alors qu'elle diminuait dans l'autre, mais la valeur totale de la première option était légèrement inférieure. A nouveau, les participants furent plus nombreux à mieux évaluer l'option croissante qu'ils ne furent à la choisir. Enfin, dans une dernière expérience inspirée de (Tversky & Griffin, 1991), les auteurs contrastèrent les deux options suivantes :

> Company A gives you a small (100 sq ft) office, and gives another employee (who has similar qualifications to you) an equally small office. Company B gives you a medium size (170 sq ft) office but gives another employee (who has similar qualifications to you) a large (240 sq ft) office.

Dans ce cas les participants prédisant une plus grande satisfaction dans la compagnie A par rapport à la compagnie B furent proportionnellement plus nombreux que ceux choisissant la compagnie A. Par la suite, Hsee et ses collègues testèrent avec succès les effets de deux autres biais. Le premier est le 'scientisme naïf', c'est-à-dire la tendance à accorder plus de poids aux traits 'durs' (objectifs, qui permettent de départager sans équivoque) qu'aux traits 'mous' (plus subjectifs, malléables). Le second le 'fonctionnalisme naïf', qui pousse les gens à accorder trop de poids aux attributs qui sont les plus directement reliés à la fonction de l'objet. Dans tous les cas, les auteurs montrent que ces biais peuvent mener les participants à prendre des décisions sous-optimales.

8.2.9 Coûts irrécupérables

Le problème des coûts irrécupérables (‘sunk cost’) est un autre ‘défaut’ de raisonnement pour lequel l’utilisation inappropriée de règles a pu être blâmée. Il s’agit de : « the tendency to continue an endeavor once an investment in money, effort, or time has been made » (Arkes & Blumer, 1985, p.124). Bien que plusieurs facteurs soient sûrement à l’origine de ce phénomène très fréquent, Arkes et Blumer ont suggéré en particulier l’utilisation inappropriée de règles s’approchant de « il ne faut pas gaspiller », ce qui selon eux expliquerait le comportement des participants face à des problèmes comme celui-ci :

> Assume that you have spent $100 on a ticket for a weekend ski trip to Michigan. Several weeks later you buy a $50 ticket for a weekend ski trip to Wisconsin. You think you will enjoy the Wisconsin ski trip more than the Michigan ski trip. As you are putting your just purchased Wisconsin ski trip ticket in your wallet you notice that the Michigan ski trip and the Wisconsin ski trip are for the same weekend? It's too late to sell either ticket, and you cannot return either one. You must use one ticket and not the other. Which ski trip will you go on? (Arkes & Blumer, 1985, p.126)

Plus de la moitié des participants choisissent alors le voyage le plus coûteux à la place de celui qu’ils apprécieraient le mieux. On peut ne pas être persuadé que la règle à laquelle les participants ont recours soit vraiment l’injonction de ne pas gaspiller. Il est possible par exemple que ce résultat soit dû à la supériorité, que nous avons déjà mentionnée, des variables quantitatives, ou même de l’argent, qui font des justifications aisées et dont il est difficile de s’échapper par peur d’être perçu comme étant incompétent (voir les divers travaux de Hsee passés en revue précédemment). Quoi qu’il en soit, il s’agirait toujours d’un effet lié au fait que la réponse qui fait tomber les participants dans l’erreur des coûts irrécupérables est plus facile à justifier. Je vais présenter plusieurs éléments qui tendent à accréditer cette hypothèse. Le premier concerne le fait que les animaux ne semblent pas être vraiment susceptibles à l’erreur des coûts irrécupérables. Nous verrons ensuite que, loin de diminuer la tendance à commettre cette erreur, le développement du

raisonnement l'augmente au contraire. Puis des travaux indiquant que la facilité avec laquelle on peut justifier les différentes options joue un rôle dans la décision seront présentés. Finalement les études montrant que les contextes qui augmentent généralement les effets des justifications (prise de décision en groupe et fait de devoir rendre des comptes) augmentent les erreurs de coûts irrécupérables seront passées en revue.

D'un point de vue adaptationniste, l'erreur des coûts irrécupérables semble vraiment étrange. On voit mal pourquoi un mécanisme d'apprentissage devrait commettre ce type d'erreur : les calculs à effectuer pour l'éviter ne semblent pas particulièrement complexes, pas plus que d'autres en tout cas. Il n'est donc guère surprenant de constater que l'erreur du Concorde ('Condorde fallacy' étant le nom que Dawkins et Brockmann [1980] ont donné à cette erreur chez les animaux) n'est pas commise par les animaux non humains. C'est en tout cas ce que concluait la revue de ces travaux faite par Arkes et Ayton en 1999. Ces auteurs avaient alors montré que les résultats précédents prétendant démontrer cette erreur chez les animaux pouvaient en fait s'expliquer par l'espoir de gains futurs. Les animaux choisissant l'option dans laquelle ils avaient commencé d'investir pouvaient raisonnablement croire que cette option s'avérerait la meilleure par la suite, ce qui n'est typiquement pas le cas dans les expériences chez les humains. Un article plus récent (Navarro & Fantino, 2005) ne mentionne pas de résultats contradictoires qui auraient été publiés depuis, mais présente par contre des expériences qui semblent, elles, démontrer un tel effet chez des pigeons.

Les oiseaux devaient, à un moment de l'expérience, faire un choix entre continuer une action déjà commencée mais risquant de n'avoir qu'un faible rendement et recommencer un nouvel essai (ce qui constituait alors la solution optimale). Une majorité d'animaux s'engagèrent alors dans le comportement sous-optimal qui correspond à l'erreur des coûts irrécupérables. Il faut néanmoins mentionner deux éléments importants. Le premier est que si la solution consistant à changer d'option (éviter l'erreur des coûts irrécupérables donc) se différencie plus nettement de l'autre, les pigeons adoptent rapidement un comportement optimal. Le second est que si un signal indique aux oiseaux le fait qu'ils se trouvent dans une situation de coûts irrécupérables, alors les animaux adoptent immédiatement un comportement optimal. Or, dans les expériences chez les humains ces derniers sont

généralement clairement informés du fait qu'ils se trouvent dans une telle situation. Donc, lorsque les pigeons se trouvent dans la situation qui se rapproche le plus de celles des participants humains, ils ne commentent pas l'erreur des coûts irrécupérables. On peut conclure de ces travaux que dans presque toutes les circonstances, et en tout cas dans celles qui se rapprochent le plus de celles auxquelles sont confrontés les participants humains, les animaux ne commettent pas l'erreur des coûts irrécupérables.

Un autre argument que l'on peut avancer pour défendre l'idée que les réponses non normatives à ces problèmes sont causées par le raisonnement et non par des mécanismes intuitifs est que les enfants en sont de plus en plus victimes. Si le raisonnement permettait de s'en défaire, on devrait s'attendre à ce qu'avec le développement, de plus en plus de participants soient capables de donner la réponse normative. Ce n'est cependant pas le cas. Arkes et Ayton (1999) mentionnent plusieurs expériences qui, s'il ne s'agit pas directement de coûts irrécupérables, en sont très proches, et qui montrent une tendance grandissante avec l'âge à donner la réponse non normative. Ainsi, des enfants avaient été confrontés à l'un des problèmes suivants (les variantes étant notées entre crochets) :

> Imagine you are at a fairground with your parents. Your mother gives you a 50 pence coin, and your father gives you a one pound coin. After walking around for a while you decide to use the 50 pence coin to buy a ticket for the merry-go-round. [But then you discover that you have lost your ticket./But then you discover that you've lost the 50 pence coin so you can't use it to buy a ticket for the merry-go-round.] Would you use the one pound coin to buy a new ticket? (Webley & Plaisier, 1997)

Lorsque c'est le ticket qui a été perdu, il est plus facile de lier ce coût irrécupérable à la décision future que lorsqu'il s'agit de la pièce, alors que cette différence ne devrait pas jouer, selon les théories économiques normatives. Le résultat fut clair. A tous les âges une grande majorité d'enfant choisit d'acheter le ticket s'ils avaient perdu la pièce. Mais si c'est le ticket qu'ils avaient perdu, seuls les plus jeunes choisirent d'en racheter un tout de même (ils étaient également 80% à le faire dans le plus jeune

groupe [5-6 ans], la proportion diminuant drastiquement avec l'âge [50% à 8-9 ans et 20% à 11-12 ans]).

Ce résultat a depuis été confirmé par des études portant spécifiquement sur l'erreur des coûts irrécupérables. Utilisant des paires de problèmes dont l'un impliquait un coût irrécupérable et l'autre non, mais qui étaient en tout points similaires par ailleurs, Klaczynski et Cottrell (2004) ont démontré qu'avec l'âge des enfants (ou plutôt des pré-adolescents, entre 8 été 14 ans) avaient de plus en plus de chances de commettre l'erreur et de choisir l'option dans laquelle des coûts irrécupérables avaient déjà été engagés. Un résultat similaire, bien que seulement présent à l'état de tendance, fut observé par Morsanyi et Handley (2008) chez des enfants se répartissant entre 5 et 11 ans. De plus, ces derniers notent chez ces enfants une corrélation positive et significative entre des tests d'habilité cognitive et la tendance à commettre l'erreur des coûts irrécupérables.

Deux séries d'études peuvent être interprétées comme montrant un rôle de la facilité de justification sur les décisions dans des situations de coûts irrécupérables. Dans deux séries d'expériences Bragger et ses collègues (J. D. Bragger, Hantula, Bragger, Kirnan, & Kutcher, 2003; J. L. Bragger, Bragger, Hantula, & Kirnan, 1998) ont étudié les effets du degré d'incertitude du feedback sur les décisions dans des situations de coûts irrécupérables. Les participants étaient engagés dans une série de décisions, et ils recevaient un feedback régulier sur les effets de ces décisions. Dans une condition, le feedback, après avoir été longtemps positif pour une action donnée, devenait assez brusquement négatif, et le restait à un taux constant. Dans l'autre condition, après un début identique, le feedback devenait en moyenne négatif (aussi négatif que dans la première condition), mais était beaucoup plus variable : il pouvait être catastrophique ou très légèrement positif par exemple. Les participants de cette seconde condition mirent beaucoup plus de temps à se désengager de l'action qui, précédemment, entraînait un feedback positif. On peut interpréter ceci facilement dans le cadre du raisonnement motivé et plus particulièrement des études de Hsee sur l'élasticité des justifications. Si, dans tous les cas, les participants sont motivés pour continuer dans l'action qui fonctionnait précédemment, ils se trouvent néanmoins aculés à un changement lorsque le feedback devient uniformément négatif : il ne leur est plus possible de justifier de ne pas changer de cap. Par contre, si le feedback est inconsistant, il leur laisse une marge : ils pourraient arguer qu'il est parfois positif,

ou qu'il n'est que légèrement négatif. De même que dans les expériences de Hsee, le fait que la moyenne du feedback soit la même dans les différentes conditions rend difficile une explication rationnelle du phénomène, et favorise donc l'explication avancée ici en termes de raisonnement motivé et d'élasticité des justifications.

Une autre série d'expériences (Soman & Cheema, 2001) c'est penchée sur le rôle des justifications dans des problèmes de coûts irrécupérables tels que celui-ci :

> You recently purchased a ticket to a rock concert by one of your favorite bands for $35. Shortly afterwards, you are invited to join a good friend on a free ski-getaway weekend. Unfortunately, the invitation is for the same weekend as the concert. The ticket is nonrefundable and non-transferable, so if you decide to go skiing, it will have to go to waste. As you stop by to pick up a paycheck of $185 (compensation for work you did at a local music store), you wonder whether you should attend the concert or go skiing.

Dans une autre condition, la fin du problème était remplacée par l'énoncé suivant :

> As you stop by to pick up a paycheck of $150 (compensation for work you did at a local music store), you were pleasantly surprised to learn that you had earned a bonus and received a total of $185. You wonder whether you should attend the concert or go skiing.

Les participants penchaient plus vers le week-end au ski dans la seconde version. Dans ce cas, il semble que la nouvelle justification fournie par la rentrée d'argent inattendue permet de compenser le 'gaspillage' de la place de concert : on pourrait dire qu'elle contrebalance la règle 'ne pas gaspiller' ou, plus généralement, le risque d'être perçu comme n'étant pas compétent.

Si le choix de l'option correspondant à l'erreur des coûts irrécupérables est ainsi guidé par le fait qu'elle est plus facile à justifier, on peut s'attendre à ce que le phénomène soit de plus grande ampleur en groupe. Comme nous l'avons vu dans le cas du cadrage, ce n'est pas une conséquence nécessaire du fait que le choix (en individuel) soit basé sur une raison, ou une règle. Dans le cas du cadrage par exemple, les réponses semblent être influencées par les justifications qui sont

rendues plus pertinentes par les différents cadres, sans que cela n'entraîne nécessairement que ces justifications soient les plus convaincantes, celles qui l'emportent en situation de groupe. Il semble néanmoins que dans le cas des erreurs de coûts irrécupérables (ou moins certaines d'entre-elles), les justifications qui les soutiennent soient bien convaincantes, car les groupes commettent encore plus l'erreur que les individus (Whyte, 1993). Enfin, on peut mentionner des expériences ayant mis les participants en situation de rendre des comptes, mais les résultats en furent variables. Alors que dans une première expérience cette manipulation tendait à renforcer l'entêtement des participants (Fox & Staw, 1979), des résultats suivants ont montré qu'elle pouvait également avoir l'effet inverse (Simonson & Nye, 1992). Ce dernier résultat pourrait être embarrassant pour l'hypothèse défendue ici. En effet, elle explique l'erreur des coûts irrécupérables par l'utilisation inappropriée de certaines règles comme justification. On pourrait alors penser que cet effet ne pourrait qu'être renforcé si les participants devaient se justifier. Cependant, les participants des expériences de Simonson et Nye étaient quelque peu différents des autres car ils avaient été, dans leurs cours, spécifiquement prévenus contre les erreurs de type coûts irrécupérables. Il est dès lors normal que les auteurs fassent, et vérifient, la prédiction suivante : « if decision makers are aware of the normative principle that sunk costs should be ignored, accountability is expected to reduce their susceptibility to the sunk cost effect (at least with hypothetical questions) » (p.420). L'analyse des résultats montra d'ailleurs que la diminution de l'erreur des coûts irrécupérables due au fait de devoir rendre des comptes n'affecta que les personnes acceptant les principes normatifs en question. Ceci ne concerne donc pas la réaction plus 'naïve' à ce type de problèmes.

Les résultats qui viennent d'être passés en revue indiquent que le raisonnement peut-être en partie, au moins, responsable des erreurs de coûts irrécupérables telles qu'elles sont généralement étudiées en prise de décision. Il est dès lors tentant d'interpréter cela comme une forme d'échec du raisonnement – un cas dans lequel les mécanismes intuitifs seraient plus efficaces. Il faut cependant faire attention avant de tirer une telle conclusion. Par exemple, les formats de présentation diffèrent considérablement entre les expériences classiques chez les humains et celles chez les animaux. Même si les problèmes partagent certaines similarités structurelles, ces modifications rendent une comparaison directe, en

termes d'efficacité, difficile. Le fait que les enfants semblent être moins victimes de ce type d'erreur est peut-être plus concluant, bien que là aussi d'autres interprétations soient envisageables. Quoiqu'il en soit, la conclusion en termes de performances du raisonnement n'est peut-être pas la plus pertinente. Ce qu'on peut constater, par contre, c'est qu'il semble fonctionner d'une façon qui s'accorde bien mieux avec les prédictions de la théorie argumentative qu'avec celles des théories classiques. Mais il s'agit là de l'argument général que va être fait dans la section suivante, n'anticipons donc pas.

8.2.10 Conclusion sur le choix basé sur des raisons

Nous avons vu dans cette partie qu'il était possible d'expliquer de nombreux résultats dans le domaine de la prise de décision grâce au 'choix basé sur des raisons'. La liste ci-dessus ne prétend pas être exhaustive, et on pourrait très bien envisager d'étendre ce cadre à d'autres effets bien connus en prise de décision, tels que l'erreur de conjonction ou la négligence des taux de base. Dans tous les cas, le choix des participants s'explique par le fait qu'il est le plus facile à justifier, celui pour lequel il est le plus aisé de trouver des raisons. Nous avions vu dans le cas des effets de cadrage que cela ne signifie pas qu'il s'agisse de la meilleure raison possible, mais simplement d'une justification qui est rendue particulièrement pertinente par le contexte. Dans certains cas, cette raison sera également la plus convaincante, et elle devrait alors avoir tendance à l'emporter dans les décisions de groupe (c'est le cas pour l'erreur des coûts irrécupérables). Mais même dans ce cas, les participants ne trouvent pas, ou rarement, la raison qu'on pourrait considérer comme *meilleure* – d'un point de vue plus normatif.

Qu'ils ne parviennent pas à cette réponse est indiqué par les deux éléments suivants. D'une part les réponses des participants dévient des prédictions des modèles normatifs de la théorie de la décision. D'autre part on peut penser que ces participants seraient convaincus de changer leurs réponses si on leur expliquait pourquoi elle ne se base pas sur les meilleures raisons, raisons qui soutiennent au contraire une autre réponse. A nouveau, cela semble être le cas pour l'erreur des coûts irrécupérables : les participants qui ont compris cette erreur et peuvent utiliser cette compréhension comme raison pour ne pas la commettre ne tombent pas dans le

piège. On peut interpréter ceci comme une démonstration de l'aspect satisficing du fonctionnement du raisonnement : il ne cherche pas la meilleure raison, simplement une raison qui, étant donné les circonstances, lui semble suffisamment convaincante. Si le raisonnement ne se satisfaisait que des meilleurs raisons, il obtiendrait alors les résultats voulus par les théories normatives[71].

Il pourrait alors être tentant de voir cet échec du raisonnement à trouver les meilleures raisons comme un simple problème de performance, lié à des contraintes cognitives – le type d'explication couramment invoqué par les théories classiques du raisonnement. Cette explication impliquerait de voir tous les résultats passés en revue ci-dessus comme des erreurs, aléatoires, du raisonnement. Mais une telle description ne correspond pas aux données : le raisonnement ne semble pas commettre n'importe quelles erreurs[72]. Il donne au contraire une réponse justifiable de façon très régulière.

Dès lors, il semble beaucoup plus économique de dire qu'il accomplit là sa fonction – et qu'il l'accomplit bien – plutôt que d'expliquer tout ces résultats comme des erreurs. On voit mal quelle contrainte pourrait être utilisée pour expliquer toutes ces erreurs : il faut accumuler des biais différents (ce qui est souvent le cas en prise de décision), chacun expliquant de façon souvent ad hoc un type de déviation par rapport aux théories normatives. On pourrait répliquer que la théorie argumentative devrait, elle, expliquer pourquoi chacune de ces justifications est rendue pertinente dans tel ou tel contexte. Il est exact qu'un tel travail sera nécessaire pour une compréhension profonde des résultats. Il n'empêche que quelque soient les raisons qui font que telle ou telle justification est utilisée, les participants répondent car ils peuvent justifier leurs réponses. Le rôle de la théorie argumentative n'est pas d'expliquer le fonctionnement de tous les mécanismes de prise de décision : ce sont principalement les mécanismes intuitifs qui déterminent la pertinence de différentes justifications, l'ordre dans lequel elles seront examinées. La théorie argumentative ne fait des prédictions que sur le fonctionnement du raisonnement. Elle prédit justement que ce dernier, s'il influence une décision, le fera dans le sens de la décision la plus facile à justifier et non d'une meilleure décision par rapport à d'autres critères. Et c'est bien ce qu'on observe.

[71] Ce qui est bien montré par le fait que les théories normatives sont justement celles qui convainquent tout le monde (ou le plus de monde en tout cas).

[72] Soulignons qu'il ne s'agit d'erreurs que par rapport à des modèles normatifs.

8.3 Conclusion sur les performances du raisonnement

Il n'y a rien de meilleur que la raison.
Cicéron, Traité sur les lois

La raison est un glaive double et dangereux.
Montaigne

La raison nous trompe plus souvent que la nature.
Vauvenargues, Réflexions et maximes

La raison est la putain du Diable.
Luther

Etant donné que le raisonnement ne mène pas forcément à de meilleures décisions, on peut se demander si les processus intuitifs ne feraient pas mieux l'affaire dans certains cas. C'est ce que pensent Dijksterhuis et ses collaborateurs, comme nous l'avons vu au début de ce chapitre. On peut en fait diviser cette question en deux : une qui concerne le niveau proximal, et une le niveau ultime.

La question précédente est généralement posée au niveau proximal, celui qui a des conséquences pratiques immédiates. Faisant face à une décision d'un certain type, est-il préférable de réfléchir longuement, de peser le pour et le contre de différentes options ainsi que le réclament les théoriciens de la prise de décision, ou vaut-il mieux au contraire n'écouter que son intuition ? Les résultats des tâches visant à comparer directement les performances de ces deux types de processus ne sont pas vraiment concluants. Mais il est tentant d'interpréter les résultats du choix basé sur des raisons dans une direction concordante avec les idées de Dijksterhuis et autres partisans de la pensée inconsciente. Si c'est le raisonnement qui nous conduit à faire ces erreurs, on ferait sûrement mieux de s'en passer, non ? Cela n'a rien d'évident. En effet, les situations typiquement présentées sont très abstraites, et les mécanismes intuitifs n'ont que peu de prise – c'est là une des raisons même qui entraîne l'activation du raisonnement. Il n'est donc pas dit qu'ils soient capables de donner des réponses consistantes, ou supérieures. Etant donné la faiblesse des

intuitions, et donc la médiocrité probable des décisions qui seraient prises sur leur base uniquement, le raisonnement ne peut pas diminuer de beaucoup les performances. Dans ces conditions, même si l'avantage du raisonnement n'était que de nous faire prendre une décision plus facilement justifiable, il est tout à fait plausible que cela compense les risques de prendre une décision légèrement inférieure.

Dans certains cas, la comparaison entre l'efficacité des mécanismes intuitifs et réflexifs est rendue plus ardue par les différences dans les formats de présentation, ou de réponse. C'est souvent le cas lorsqu'on souhaite comparer les performances des humains et d'autres animaux[73]. Bien que certains problèmes utilisés chez les hommes et les autres animaux partagent des similarités structurelles, ces différences rendent une comparaison directe difficile. Même lorsque les problèmes sont aussi soigneusement appariés que possible, certains facteurs sont presque impossible à égaliser (tels que la motivation : les comités d'éthiques tendent à faire grise mine si on leur propose d'assoiffer les participants…).

Il est également difficile de comparer les performances des mécanismes intuitifs et réflexifs chez des participants. Ainsi, Maloney et ses collègues ont construit des tâches qui sont structurellement analogues à certaines tâches de prise de décision, mais qui engagent la coordination visuo-motrice (Maloney, Trommershauser, & Landy, 2007; Trommershauser, Landy, & Maloney, 2006; Trommershauser, Maloney, & Landy, 2008). Les participants atteignent alors généralement des performances optimales, alors même qu'elles sont mauvaises dans les tâches plus classiques de prise de décision. Mais des différences importantes subsistent entre les tâches, telle que la quantité d'essais. Ces difficultés ne signifient cependant pas qu'on ne peut tirer aucune conclusion intéressante de ces travaux, ce qui nous conduit au versant ultime de la question.

Au niveau ultime la question devient la suivante : est-ce que les problèmes que le raisonnement est censé avoir évolué pour résoudre auraient pu être mieux résolus par des mécanismes intuitifs – nouveaux si besoin ? C'est peut-être pour répondre à cette question que les tâches structurellement analogues aux tâches de

[73] Voir par exemple Arkes et Ayton (1999) dans le cas de l'erreur des coûts irrécupérables, Fantino, Kanevsky et Charlton (2005) pour la négligence des taux de base, Chen, Lakshminarayanan et Santos (2006) pour l'aversion aux pertes, Rosati, Stevens, Hare et Hauser (2007) pour l'actualisation temporelle ('temporal discounting'), et Egan, Santos et Bloom (2007) pour la dissonance cognitive.

prise de décision classiques, mais aménagées pour pouvoir être résolues par des mécanismes intuitifs sont les plus intéressantes. Si les analogies entre les tâches sont suffisamment profondes, et que les mécanismes intuitifs parviennent à les résoudre, on peut raisonnablement penser que des mécanismes intuitifs qui auraient évolué pour s'adapter aux problèmes plus classiques pourraient parfaitement les résoudre. Il s'agit donc de preuves expérimentales en faveur d'un argument que l'on peut par ailleurs faire sur des considérations purement théoriques ayant trait à la complexité des tâches. Bien que la complexité soit très dure à quantifier, il semble difficile de soutenir que les tâches utilisées en prise de décision soient plus complexes que celles que résolvent constamment nos systèmes perceptifs, moteurs, ou de prise de décision intuitifs. Si cela ne veut pas dire que, en l'état actuel des choses, il est préférable d'utiliser des mécanismes intuitifs, cela signifie que s'il y avait eu de réelles pressions de sélection pour résoudre ce type de problème, on voit mal pourquoi de nouveaux mécanismes intuitifs – ou d'anciens 'bricolés' – n'auraient pu évoluer pour les résoudre. Cette conclusion conforte donc l'idée que le raisonnement n'a pas pour fonction de résoudre ce type de problème.

Conclusion

L'objectif de cette thèse a été de montrer que la théorie argumentative du raisonnement est à la fois plausible d'un point de vue évolutionniste et peut rendre compte de nombreux résultats qui, sinon, requièrent des explications différentes et souvent ad hoc.

On peut diviser les arguments empiriques en deux grandes catégories : négatifs et positifs. Les arguments négatifs portent sur des erreurs du raisonnement, ou plutôt des déviations par rapport aux théories normatives. Ce sont les nombreux cas dans lesquels le raisonnement ne fait pas ce qu'il devrait faire à en croire les théories classiques. Il ne s'agit là que de soutien indirect pour la théorie argumentative. En fait, il s'agit plutôt d'arguments contre les théories classiques, dont la théorie argumentative ne bénéficie que parce qu'elle est une des seules concurrentes à ces théories. Les arguments positifs correspondent aux cas dans lesquels le raisonnement se comporte d'une façon prédite par la théorie argumentative.

Les premiers arguments avancés étaient positifs : le raisonnement accompli bien ce qui est, selon la théorie argumentative, sa fonction primaire : évaluer et produire des arguments. Les résultats en faveur de cette conclusion viennent de plusieurs disciplines. En psychologie sociale, la tradition de recherche sur la persuasion et le changement d'attitude a établi que lorsque les participants sont motivés ils sont parfaitement à même d'être sélectivement influencés par de bons arguments. Au croisement de la psychologie sociale et de la psychologie du raisonnement, l'étude du raisonnement en groupe montre que les membres d'un groupe discutant d'un problème de raisonnement convergent vers la meilleure réponse présente dans le groupe. Cela signifie à la fois que les personnes ayant compris la bonne réponse sont capables de trouver de bons arguments pour en convaincre les autres, et que ces derniers peuvent discriminer les bons arguments des mauvais. Enfin, les études s'étant portées plus spécifiquement sur nos capacités d'argumentation tendent à montrer que les participants ont de bonnes performances. Lorsque ce n'est pas le cas, les limites des participants s'expliquent bien dans le cadre de la théorie argumentative. Ainsi, le fait que les participants ne trouvent pas

spontanément de contre-arguments visant leurs propres idées, s'il peut être considéré comme problématique d'un point de vue normatif, est exactement ce qu'on attend d'un mécanisme visant à persuader.

A l'exception de ces derniers résultats qui soutiennent plus directement la théorie argumentative, on pourrait très bien dire que ceux qui viennent d'être mentionnés ne permettent pas de la départager d'autres théories du raisonnement. En effet, rien dans ces théories ne dit que le raisonnement devrait être inefficace en situation d'argumentation. Ces données sont néanmoins importantes pour deux raisons. La première est que si les gens étaient réellement mauvais en argumentation, il s'agirait d'un coup fatal pour la théorie défendue ici. La seconde est que les bonnes performances observées dans le cadre de l'argumentation créent un contraste marqué avec les mauvaises performances dans des problèmes de psychologie du raisonnement. Et ce alors même que ces problèmes sont soit similaires, soit plus simples par bien des aspects que les problèmes étudiés dans le cadre de l'argumentation.

Les premiers arguments que l'on peut tirer de la psychologie du raisonnement sont négatifs : dans les problèmes abstraits classiques, les performances sont souvent catastrophiques. Pour reprendre un argument avancé plus haut, si les mécanismes de raisonnement avaient vraiment pour fonction de faire des déductions il s'agirait d'un échec de conception inédit dans les annales de l'évolution. C'est là un argument pesant directement contre les théories pour lesquelles le raisonnement est une adaptation devant remplir ce type de fonction. Les différents contre-arguments que l'on peut présenter contre cette critique ont été examinés dans la première partie, et aucun n'est vraiment convaincant.

Mais l'examen de résultats plus précis dans ce même domaine a également permis de rassembler un faisceau d'arguments positifs pour la théorie argumentative. Les problèmes de psychologie du raisonnement ne se placent pas, en grande majorité, dans un contexte argumentatif naturel : personne n'essaie de convaincre les participants de quoi que ce soit. Par contre, les participants ont de fortes chances d'utiliser le raisonnement pour chercher à justifier leurs réponses obtenues principalement sur la base de mécanismes intuitifs. C'est bien ce qui semble se passer, que ça soit dans des tâches de test d'hypothèse, de raisonnement conditionnel ou de raisonnement syllogistique. Dans ces tâches, des mécanismes intuitifs, qu'il

s'agisse de mécanismes de pertinence conversationnelle ou de vérification de cohérence par exemple, fournissent une réponse au participant, réponse qu'il s'efforcera ensuite de justifier. A aucun moment les participants n'essaieront de falsifier une réponse qu'ils pensent intuitivement être la bonne. S'ils s'engagent bien parfois dans de la falsification, c'est vis-à-vis d'éléments qu'ils pensent justement être faux et qu'ils veulent réfuter. Il peut s'agir d'hypothèses proposées par d'autres personnes (dans le 2,4,6), de règles qu'ils sont motivés pour rejeter (dans la tâche de sélection de Wason) ou de conclusions contraires à leurs représentations ou leurs croyances (dans les syllogismes).

Ces observations peuvent être interprétées comme des arguments négatifs : étant donné que la falsification est souvent nécessaire pour que le raisonnement se conforme aux théories normatives, on peut se demander pourquoi les participants ne s'y engagent pas plus spontanément. Mais pour la théorie argumentative, c'est précisément cela que l'on attend : le raisonnement des participants devrait être biaisé, il ne devrait servir qu'à soutenir une position qu'ils sont motivés pour défendre. Il s'agit donc bien également d'un argument positif, car on voit mal comment expliquer ces résultats simplement sur la base de limites cognitives ou d'autres contraintes.

Une série de travaux se situant principalement dans le domaine de la psychologie sociale est parvenue à des conclusions similaires : les études sur le raisonnement motivé. On retrouve sous cet ombrelle des travaux assez différents, mais qui convergent tous vers la conclusion suivante. Si les participants sont motivés pour parvenir à une certaine conclusion et qu'ils utilisent le raisonnement, son fonctionnement sera fortement biaisé et il ne cherchera que des arguments soutenant cette conclusion. Cependant, la disponibilité de justifications va également influencer la réponse : plus il sera facile de trouver des justifications pour défendre la position qui motive les participants, plus ceux-ci auront en effet des chances de la prendre. Il s'agit là d'arguments positifs soutenant la théorie argumentative : lorsque nous sommes motivés pour défendre une position, on attend précisément ce type de fonctionnement du raisonnement.

Mais on peut également considérer ces résultats comme des arguments négatifs : dans ces conditions, le raisonnement échoue à 'corriger' les motivations des participants, il n'est pas utilisé pour parvenir à une meilleure réponse, mais uniquement pour se justifier. Ceci est d'autant plus vrai que ces mêmes biais peuvent

très bien toucher des participants qui pensent être motivés pour donner une réponse correcte, des participants qui n'ont pas du tout l'impression de raisonner de façon biaisée. Il s'agit donc d'une autre occasion pour laquelle le raisonnement échoue à faire ce qu'il devrait faire pour les théories classiques : aider à prendre de meilleures décisions. Ainsi, le raisonnement motivé peut mener à l'évaluation biaisée d'arguments, qui à son tour peut entraîner une polarisation des attitudes qui serait bien difficile à justifier normativement. Même en l'absence d'arguments à évaluer, le simple fait de penser à un sujet peut pousser les gens vers des attitudes plus extrêmes ou renforcer leur confiance sans qu'il y ait pour cela de bonnes raisons. Il est également possible que les gens maintiennent, à cause de ce type de biais, des croyances 'périmées', qui devraient être révisées suite à l'acquisition de nouvelles données.

Toujours en psychologie sociale, mais dans un champ différent, des chercheurs ont analysé les performances du raisonnement face à divers problèmes de jugement ou de prise de décision, qu'il s'agisse de choisir le meilleur item dans une liste, de prédire le résultat de matchs de basket, ou même de prédire son propre comportement. A travers de nombreuses expériences, les résultats furent souvent décevants pour le raisonnement, n'aidant pas dans certains cas, entraînant une baisse de performance dans d'autres. Il ne s'agit pas de dire qu'il ne faut jamais réfléchir avant de prendre une décision, mais ces données montrent que le raisonnement n'est pas une panacée, même face à des problèmes qui sont typiquement ceux pour lesquels les théories classiques du raisonnement comme de la prise de décision prescriraient son utilisation. Il ne s'agit là cependant que d'un argument négatif.

De nombreux résultats dans le domaine de la prise de décision, s'ils renforcent cet argument négatif, offrent également un soutien positif pour la théorie argumentative : ils se situent dans le cadre du choix basé sur des raisons. Les chercheurs prenant cette perspective ont montré que lorsque les gens font des choix qui ne s'accordent pas avec les normes des théories de la décision, ils le font souvent car leurs choix sont les plus faciles à justifier, ceux pour lesquels ont peut fournir des raisons. Il s'agit bien là d'un argument négatif car les choix vers lesquels le raisonnement pousse ne sont pas optimaux. Mais il s'agit surtout d'un argument positif : le raisonnement ne fait pas que commettre des erreurs, il agit de la façon

prédite par la théorie argumentative en poussant vers les choix les plus facilement justifiables.

De tous les cas qui viennent d'être examinés, on peut conclure que le raisonnement fait mal ce qu'il est censé faire selon les théories classiques (arguments négatifs), mais fait bien ce qui est prédit par la théorie argumentative (arguments positifs). Le raisonnement semble bien fonctionner comme un mécanisme de production et d'évaluation d'arguments.

Même en acceptant cette conclusion, certains pourraient s'interroger sur l'utilité de ce type de considérations, portant principalement sur des questions ultimes, sur la fonction du raisonnement. On peut raisonnablement espérer, cependant, qu'une meilleure connaissance de la fonction du raisonnement permette de faire des prédictions testables sur la façon dont il fonctionne : il s'agit là de la 'pensée adaptative' chère aux psychologues évolutionnistes. C'est par exemple ce que j'ai essayé de montrer dans le chapitre 2. Partant des tâches que le raisonnement doit accomplir (principalement évaluer des arguments), j'ai essayé d'en inférer certaines caractéristiques de son fonctionnement. Ceci permet de faire des prédictions sur le fonctionnement du raisonnement. Des moyens de tester directement ces prédictions vont être proposés dans la prochaine section.

Considérations expérimentales

Avant de suggérer quelques pistes possibles de travaux expérimentaux, il faut noter que la théorie argumentative, seule, ne peut guère faire de prédictions sur les résultats précis d'expériences, et ce même s'il s'agit d'expériences de raisonnement 'pures'. La recherche de contextes abstraits pour les tâches de raisonnement, qui avait pour objectif de n'impliquer que le minimum de capacités ou de connaissances en dehors du raisonnement, a eu pour effet d'ôter toute motivation naturelle pour raisonner sur le matériel lui-même. A la place, les participants tentent de justifier une conclusion proposée par des mécanismes intuitifs. Ce sont ces mécanismes qui décident, en premier lieu, de la réponse des participants. Si les mécanismes intuitifs jouent un rôle déterminant même dans ces tâches, cela signifie qu'il est en fait

impossible de créer une tâche qui isolerait totalement les mécanismes de raisonnement. Une conséquence inévitable est que pour prédire les résultats à une tâche qui pourrait impliquer le raisonnement, il faut tout d'abord pouvoir faire des prédictions sur les réponses qui seront favorisées par les mécanismes intuitifs.

Bien que ceci puisse sembler être une limitation, la place très importante laissée aux mécanismes intuitifs est au contraire une des forces de la théorie argumentative. Elle se trouve en effet dans une position privilégiée pour expliquer les effets conversationnels, de contenu, de contexte, des croyances préalables, etc. qui tendent généralement à être vus comme des imperfections du raisonnement. Dans le cadre de la présente théorie, il ne s'agit que d'effets tout à fait normaux, et adaptatifs, des mécanismes intuitifs.

Le rôle nécessaire des mécanismes intuitifs ne signifie pas non plus que la théorie argumentative ne peut pas faire de prédictions du tout. D'une part elle peut faire des prédictions générales sur le fonctionnement du raisonnement, prédictions confirmées par les travaux passés en revue dans cette thèse. D'autre part elle peut faire des prédictions spécifiques, pourvu que l'on spécifie le fonctionnement des mécanismes intuitifs risquant de jouer un rôle dans la tâche. De plus, elle fait des prédictions qui lui sont souvent propres sur l'effet que peuvent avoir certains contextes ou certaines instructions. Par exemple, le besoin de se justifier devrait renforcer les effets du raisonnement. Le contexte social dans lequel se trouvent les participants devrait influencer ce qu'ils considèrent comme étant de bonnes justifications. On trouve des résultats concordants en psychologie sociale comme en prise de décision, et il sera intéressant de poursuivre ces recherches. On pourra étendre les études de raisonnement en groupe aux autres problèmes classiques de la psychologie du raisonnement. Utiliser ces mêmes problèmes en contexte argumentatif, afin que les participants soient naturellement motivés pour les examiner à l'aide du raisonnement. Ou, dans l'autre sens, analyser plus finement les arguments utilisés dans les études sur la persuasion et le changement d'attitude. Se servir des méthodes du choix basé sur des raisons pour rendre compte d'autres problèmes de prise de décision, et peut-être de raisonnement. En rassemblant sous un même modèle explicatif des résultats de diverses disciplines, la théorie argumentative pourrait faciliter l'échange de méthodologies qui étaient jusqu'à présent l'apanage de ces différents domaines.

Enfin, si la théorie argumentative met l'accent sur le rôle des mécanismes intuitifs, elle prédit également que le raisonnement jouera un rôle dans des situations ou on ne l'attend pas forcément. Les situations expérimentales tendent à déclencher le raisonnement : sachant que quelqu'un va s'intéresser à leurs réponses, les participants souhaitent s'assurer qu'elles seront justifiables. En plus de ces motivations externes, il arrive souvent que les intuitions des participants quant à la réponse correcte soient faibles ou conflictuelles. Ceci est principalement dû à un biais dans la façon dont sont construites les expériences : étant donné que les mécanismes intuitifs, lorsqu'ils sont normalement activés, tendent à fournir des réponses correctes qu'on considérera souvent comme étant triviales, les expérimentateurs vont essayer de s'éloigner de ces effets plafond en rendant les tâches plus complexes, en brouillant les pistes. Ces manipulations auront pour effet de rendre les intuitions moins claires, plus faibles ou plus conflictuelles, et donc d'augmenter les chances que le raisonnement intervienne. Dans certaines expériences ciblant les processus intuitifs, le raisonnement jouera donc un rôle, à l'insu parfois des expérimentateurs. C'est un des problèmes qui se pose par exemple lorsque l'on souhaite comparer les performances d'humains et d'autres animaux : même si les modes de présentation et de réponse sont très proches, il est toujours possible que les réponses des humains soient influencées par le raisonnement, et ne reflètent donc qu'imparfaitement le fonctionnement des mécanismes intuitifs.

D'autres soutiens

La théorie argumentative pourrait tirer de nouveaux soutiens dans plusieurs domaines. La psychologie sociale constitue un domaine d'extension privilégié. J'ai déjà utilisé plusieurs résultats tirés de ce domaine, mais il ne s'agit que ceux dont les applications étaient les plus directes. Les théories à processus duel, originellement construites pour rendre compte de résultats en persuasion et changement d'attitude, s'étendent maintenant à toute la psychologie sociale, de l'influence des stéréotypes (Bodenhausen, Macrae, & Sherman, 1999), en passant par la perception des personnes (Uleman, 1999), jusqu'à ce concept central que sont les attitudes (Wilson et al., 2000). De plus, les chercheurs de ce domaine sont de plus en plus nombreux à faire un lien direct avec les théories à processus duel en psychologie du

raisonnement, et à tester certaines des implications de ces théories dans le domaine social (Deutsch, Gawronski, & Strack, 2006; Gawronski & Bodenhausen, 2006). Etant donné que de nombreuses tâches de psychologie sociale contiennent les ingrédients qui devraient déclencher le raisonnement (intuitions faibles ou conflictuelles, motivation pour ne pas paraître incompétent ou malveillant), on peut s'attendre à ce qu'une 'couche' de raisonnement se superpose aux mécanismes intuitifs portant sur le domaine social. Pour ne prendre qu'un exemple, une des théories dominantes de l'usage des stéréotypes explique leur expression par une interaction entre des intuitions et des mécanismes qui orientent les participants vers des réponses qu'ils peuvent justifier (Crandall & Eshleman, 2003).

Toujours en psychologie sociale, deux domaines pourraient être riches de données soutenant la théorie argumentative. Rappelons qu'elle prédit que des intuitions conflictuelles devraient favoriser le déclenchement du raisonnement. Deux sources de travaux s'ouvrent alors. D'une part les recherches sur les impressions métacognitives : comment des impressions de fluence – ou au contraire de manque de fluence – peuvent déclencher des mécanismes de raisonnement (voir par exemple Oppenheimer, 2008; Whittlesea & Williams, 2001a, 2001b, et également N. Schwarz, 2004, sur les impressions métacognitives plus généralement). D'autre part, et surtout, l'immense domaine de la dissonance cognitive. L'objectif serait alors de montrer (i) que le raisonnement est bien activé lorsque la théorie argumentative le prédit et (ii) qu'il fonctionne alors de la façon dont elle le prédit. Les résultats des recherches sur la dissonance cognitive se prêteraient particulièrement bien à ces interprétations. Même si, du point de vue proximal, la recherche de consonance peut être vue comme visant le maintien d'une bonne vision de soi-même (Cooper, 2007), il semble que d'un point de vue ultime ceci ne peut guère s'expliquer que par le besoin d'apparaître compétent ou bienveillant aux yeux des autres. Mais cet argument reste à développer.

A la frontière entre la psychologie sociale et la psychologie morale, de nombreux travaux accréditent l'hypothèse que le raisonnement se cantonne souvent à jouer un rôle de recherche de rationalisations ou de justifications dans les décisions morales (voir par exemple Haidt, 2001). Plus récemment, des modèles s'inspirant directement des théories à processus duel ont été présentés (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2001). Certains résultats se situent dans la lignée de ceux sur le

raisonnement motivé : lorsque les gens sont motivés pour prendre une décision qui peut être moralement répréhensible, le raisonnement leur sert à trouver des justifications, et il augmente donc les chances que les personnes s'engagent bien dans cette action (Batson, Thompson, Seuferling, Whitney, & Strongman, 1999; Bersoff, 1999; Dana, Weber, & Kuang, 2007; Mazar, Amir, & Ariely, In prep; Valdesolo & DeSteno, 2007, 2008).

Un autre domaine d'ouverture possible est celui de la psychologie du développement. J'ai essayé de montrer chez les adultes l'existence d'un contraste fort entre les bonnes capacités en argumentation et les mauvaises performances dans des tâches de raisonnement abstrait. Il semble qu'on puisse trouver un contraste similaire – plus marqué peut-être encore – chez de très jeunes enfants. Les conclusions des travaux de Nancy Stein et ses collègues sur le développement de l'argumentation sont particulièrement frappantes (Stein & Albro, 2001; Stein & Bernas, 1999; Stein & Miller, 1993). Sur la base de l'analyse de corpus et d'expériences, ils « montrent que mêmes les enfants les plus jeunes [3 ans] prenant part à des débats ['arguments'] sont capables de générer et de penser à des raisons positives et négatives pour s'engager dans des actions différentes ou pour avoir telle ou telle croyance » (Stein & Bernas, 1999). De même que chez les adultes la motivation joue un rôle essentiel pour parvenir à ces performances. Les auteurs expliquent les très bonnes performances de très jeunes enfants par le fait que les situations qu'ils ont étudiées sont « personnellement importantes pour les jeunes enfants, et ont un effet direct sur leurs buts, leurs croyances, leur bien-être ». (Ibid.). On peut également rappeler ici les nombreux résultats démontrant l'efficacité du raisonnement en groupe chez les enfants (voir sections 5.1.1 et 5.2.2). De plus, le développement des capacités d'argumentation ne s'accompagne pas d'une diminution des biais : les « participants aux conflits ['arguers'] de toutes les classes d'âge, des élèves de maternelle aux adultes […] montrent les mêmes biais dans leur compréhension et leur remémoration des conflits, et ce indépendamment de leur âge » (Stein & Albro, 2001, p.130). Il semble donc qu'on observe chez les enfants un pattern similaire à celui observé chez les adultes, et également en faveur de la théorie argumentative.

Ouverture

La vision du raisonnement qui vient d'être présentée pourrait sembler exagérément pessimiste : le raisonnement est fondamentalement biaisé, la recherche de l'objectivité est illusoire. Si c'est peut-être le cas au niveau individuel, ça ne l'est pas nécessairement au niveau collectif. Lorsque d'autres personnes sont là pour défendre des points de vue différents, pour attaquer nos préconceptions, le raisonnement donne le meilleur de lui-même. Alors, loin d'être des limites, nos biais ne sont que des outils d'une division du travail qui permet au collectif d'atteindre plus facilement son objectif.

Mais au-delà de l'efficacité ainsi acquise, le fait que le raisonnement n'est nulle part plus à son aise que durant un débat fait aussi partie du plaisir qu'on prend à l'utiliser. Pour ceux d'entre nous dont raisonner est une des activités principales, il serait bien triste qu'il ne fonctionne correctement qu'en solitaire. Spinoza lui-même, pourtant pas le plus social des êtres, l'a bien reconnu : « Rien ne peut être plus utile à l'homme pour conserver son être et jouir de la vie raisonnable que l'homme lui-même quand la raison le conduit » (Spinoza, Ethique IV, appendice IX).

Bibliographie

Achen, C. H., & Bartels, L. M. (2006). It Feels Like We're Thinking: The Rationalizing Voter and Electoral Democracy, *Annual Meeting of the American Political Science Association, Philadelphia, August*.

Acker, F. (2008). New findings on unconscious versus conscious thought in decision making: additional empirical data and meta-analysis. *Judgment and Decision Making, 3*(4), 292-303.

Allport, F. (1924). *Social Psychology*. Boston: Houghton Mifflin.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*(6), 1037-1049.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409-429.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14*, 471-517.

Anderson, T., Howe, C., Soden, R., Halliday, J., & Low, J. (2001). Peer interaction and the learning of critical thinking skills in further education students. *Instructional Science, 29*(1), 1-32.

Anderson, T., Howe, C., & Tolmie, A. (1996). Interaction and mental models of physics phenomena: Evidence from dialogues between learners. In J. Oakhill & A. Garnham (Eds.), *Mental Models in Cognitive Science: Essays in Honour of Phil Johnson-Laird* (pp. 247-273). Hove: The Psychology Press.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (In Press). Large Stakes and Big Mistakes.

Arkes, H. R., & Ayton, P. (1999). The sunk cost and Concorde effects: Are humans less rational than lower animals. *Psychological Bulletin, 125*(5), 591–600.

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes, 35*(1), 124-140.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes, 37*(1), 93-110.

Bacon, F. (1620). *Novum Organum*.

Bailenson, J. N., & Rips, L. J. (1996). Informal reasoning and burden of proof. *Applied Cognitive Psychology, 10*(7), 3-16.

Bara, B. G., Bucciarelli, M., & Johnson-Laird, P. N. (1995). Development of syllogistic reasoning. *American Journal of Psychology, 108*, 157-193.

Barber, B. M., Heath, C., & Odean, T. (2003). Good Reasons Sell: Reason-Based Choice Among Group and Individual Investors in the Stock Market. *Management Science, 49*(12), 1636-1652.

Bargh, J. A. (2002). Losing Consciousness: Automatic Influences on Consumer Judgment, Behavior, and Motivation. *Journal of Consumer Research, 29*(2), 280-285.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230-244.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning, 1*, 221-235.

Barrouillet, P., Grosset, N., & Lecas, J. F. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition, 75*(3), 237-266.

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition, 11*(3), 211-227.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77*(3), 525-537.

Baum, L. A., Danovitch, J. H., & Keil, F. C. (2007). Children's sensitivity to circular explanations. *Journal of Experimental Child Psychology, 100*(2), 146-155.

Bazerman, M. H., Loewenstein, G. F., & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly, 37*(2).220-240.

Bem, D. J., Wallach, M. A., & Kogan, N. (1965). Group decision making under risk of aversive consequences. *Journal of Personality and Social Psychology, 95*, 453-460.

Berridge, K. C., & Winkiehnan, P. (2003). What is an unconscious emotion?(The case for unconscious" liking"). *Cognition and Emotion, 17* (2), 181-211.

Berry, D. C., & Dienes, Z. (1993). *Implicit learning*. Hove: Erlbaum.

Bersoff, D. M. (1999). Why Good People Sometimes Do Bad Things: Motivated Reasoning and Unethical Behavior. *Personality and Social Psychology Bulletin, 25*(1), 28.

Billig, M. (1996). *Arguing and Thinking: A Rhetorical Approach to Social Psychology*. Cambridge: Cambridge University Press.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal Reasoning in Rats. *Science, 311*(5763), 1020-1022.

Blanchette, I., & Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. *Memory & Cognition, 28*(1), 108-124.

Blanchette, I., & Dunbar, K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory & Cognition, 29*(5), 730-735.

Blinder, A. S., & Morgan, J. (2000). Are two heads better than one?: An experimental analysis of group vs. individual decision making. *NBER Working Paper*.

Blum-Kulka, S., Blondheim, M., & Hacohen, G. (2002). Traditions of dispute: from negotiations of talmudic texts to the arena of political discourse in the media. *Journal of Pragmatics, 34*(10-11), 1569-1594.

Bodenhausen, G. V., Macrae, C. N., & Sherman, J. W. (1999). On the dialectics of discrimination: Dual processes in social stereotyping. In S. Chaiken & Y. Trope (Eds.), *Dual-Process Theories in Social Psychology*. New York: The Guilford Press.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation, 19*(2), 442-477.

Boiney, L. G., Kennedy, J., & Nye, P. (1997). Instrumental Bias in Motivated Reasoning: More When More Is Needed. *Organizational Behavior and Human Decision Processes, 72*(1), 1-24.

Boinski, S., Moraes, E., Kleiman, D. G., Dietz, J. M., & Baker, A. J. (1994). Intra-group vocal behavior in wild golden lion tamarins, Leontopithecus rosalia: Honest communication of individual activity. *Behaviour 130*, 53-75.

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision making and performance. *Organizational Behavior and Human Decision Processes, 88*, 719–736.

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A Review of the Effects of Financial Incentives on Performance in Laboratory Tasks: Implications for Management Accounting. *Journal of Management Accounting Research, 12*(1), 19-64.

Bonner, S. E., & Sprinkle, G. B. (2002). The Effects of Monetary Incentives on Effort and Task Performance: Theories, Evidence, and a Framework for Research. *Accounting, Organizations and Society, 27*(4-5), 303-345.

Bovet, D. (in prep). Ethologie et évolution. In J.-B. Van der Henst & H. Mercier (Eds.), *Evolution et cognition*. Grenoble: Presses Universitaires de Grenoble.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *PNAS, 105*(38), 14325-14329.

Bragger, J. D., Hantula, D. A., Bragger, D., Kirnan, J., & Kutcher, E. (2003). When success breeds failure: history, hysteresis, and delayed exit decisions. *Journal of Applied Psychology, 88*(1), 6-14.

Bragger, J. L., Bragger, D. H., Hantula, D. A., & Kirnan, J. P. (1998). Hysteresis and uncertainty: The effect of information on delays to exit decisions. *Organizational Behavior and Human Decision Processes, 74*, 229-253.

Brannon, L. A., Tagler, M. J., & Eagly, A. H. (2007). The moderating role of attitude strength in selective exposure to information. *Journal of Experimental Social Psychology, 43*(4), 611-617.

Brem, S. K. (2003). Structure and pragmatics in informal argument: Circularity and question-begging. *Trends in Cognitive Sciences, 7*(4), 147-149.

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24*, 573–604.

Briley, D. A., Morris, M. W., & Simonson, I. (2000). Reasons as carriers of culture: Dynamic versus dispositional models of cultural influence on decision making. *Journal of Consumer Research, 27*(2), 157-178.

Brock, T. C. (1967). Communication discrepancy and intent to persuade as determinants of counterargument production. *Journal of Experimental Social Psychology, 3*(3), 269-309.

Brogan, H. (2006). *Alexis de Tocqueville: A Life*. New Haven: Yale University Press.

Brown, C. L., & Carpenter, G. S. (2000). Why is the trivial important? A reasons-based account for the effects of trivial attributes on choice. *Journal of Consumer Research, 26*(4), 372-385.

Burt, A., & Trivers, R. (2006). *Genes in Conflict: The Biology of Selfish Genetic Elements*. Cambridge, MA: Harvard University Press.

Byrne, R. W., & Whiten, A. (Eds.). (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. New York: Oxford University Press.

Cacioppo, J. T., & Petty, R. E. (1979). Effects of message repetition and position on cognitive response, recall, and persuasion. *Journal of Personality and Social Psychology, 37*(1), 97-109.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131.

Camerer, C., & Hogarth, R. M. (1999). The effect of financial incentives on performance in experiments: a review and capital-labor theory. *Journal of Risk and Uncertainty, 19*, 7-42.

Cameron, J. A., & Trope, Y. (2004). Stereotype-Biased Search and Processing of Information About Group Members. *Social Cognition, 22*(6), 650-672.

Campan, R., & Scapani, F. (2002). *Ethologie*. Bruxelles: De Boeck.

Carpenter, G. S., Glazer, R., & Nakamoto, K. (1994). Meaningful Brand from Meaningless Differentiation: The Dependence on Irrelevant Attributes. *Journal of Marketing Research, 31*(3), 339-350.

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212-252). New York: Guilford Press.

Chaiken, S., & Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. New York: The Guilford Press.

Chaiken, S., & Yates, S. (1985). Affective-cognitive consistency and thought-induced attitude polarization. *Journal of Personality and Social Psychology, 49*(6), 1470-1481.

Cheney, D. L., & Seyfarth, R. M. (1990). *How Monkeys See the World*. Chicago: Chicago University Press.

Chernev, A. (2005). Context effects without a context: Attribute balance as a reason for choice. *Journal of Consumer Research, 32*(2), 213-223.

Chomsky, N., & Barsamian, D. (2002). *De la propagande.* Paris: Librairie Artheme Fayard.

Cicéron. (52 avant JC). *Traité des lois*.

Claparède, E. (1923). Préface In *Le Langage et la pensée chez l'enfant, Jean Piaget*. Neuchâtel: Delachaux et Niestlé.

Clayton, N. S., & Emery, N. J. (2007). The social life of corvids. *Current Biology, 17*(6), R652-656.

Clément, F., Koenig, M. A., & Harris, P. (2004). The ontogeny of trust. *Mind and Language, 19*(4), 360-379.

Conlon, E. J., & Wolf, G. (1980). The moderating effects of strategy, visibility, and involvement on allocation behavior: An extension of Staw's escalation paradigm. *Organizational Behavior and Human Performance, 26*, 172-192.

Cooper, J. (2007). *Cognitive Dissonance: Fifty Years of a Classic Theory*. London: Sage.

Corner, A., Hahn, U., & Oakfsord, M. (2006). The slippery slope argument: probability, utility and category reappraisal. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*.

Cowley, M., & Byrne, R. M. J. (2005). *When falsification is the only path to truth.* Paper presented at the Twenty-Seventh Annual Conference of the Cognitive Science Society, Stresa, Italy.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129*(3), 414-446.

Crano, W. D., & Prislin, R. (2006). Attitudes and persuasion. *Annual Review of Psychology, 57*, 345-374.

Croson, R. T. A. (1999). The disjunction effect and reason-based choice in games. *Organizational Behavior and Human Decision Processes, 80*(2), 118-133.

Dall, S. R. X., Giraldeau, L. A., Olsson, O., McNamara, J. M., & Stephens, D. W. (2005). Information and its use by animals in evolutionary ecology. *Trends in Ecology & Evolution, 20*(4), 187-193.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*(1), 67-80.

Davis, J. H. (1973). Group decisions and social interactions: A theory of social decision schemes. *Psychological Review, 80*, 97-125.

Dawkins, R. (1990). *Le Gène égoïste*. Paris: Armand Colin.

Dawkins, R., & Brockmann, H. J. (1980). Do digger wasps commit the Concorde fallacy. *Animal Behaviour, 28*, 892-896.

Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (pp. 282-309). Oxford: Basil Blackwell Scientific Publications.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated Reasoning and Performance on the was on Selection Task. *Personality and Social Psychology Bulletin, 28*(10), 1379.

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little, Brown.

Dessalles, J.-L. (2000). *Aux Origines du langage*. Paris: Hermes Science Publications.

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the Boundaries of Automaticity: Negation as Reflective Operation. *Journal of Personality and Social Psychology, 91*(3), 385.

Dijksterhuis, A. (2004). Think different: the merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology, 87*(5), 586-598.

Dijksterhuis, A., Aarts, H., & Smith, P. K. (2002). The Power of the Subliminal: On Subliminal Persuasion and Other Potential Applications. In R. Hassin, J. Uleman & B. J. (Eds.), *The New Unconscious*. New York: Oxford University Press.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception-behavior expressway. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 33, pp. 1-40). San Diego, CA: Academic Press.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science, 311*(5763), 1005-1007.

Dijksterhuis, A., & van Olden, Z. (2006). On the benefits of thinking unconsciously: Unconscious thought can increase post-choice satisfaction. *Journal of Experimental Social Psychology, 42*(5), 627-631.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63*(4), 568-584.

Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A., & Lockhart, L. K. (2003). Spontaneous Skepticism: The Interplay of Motivation and Expectation in Responses to Favorable and Unfavorable Medical Diagnoses. *Personality and Social Psychology Bulletin, 29*(9), 1120.

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology, 75*(1), 53-69.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks, 12*(7-8), 961-974.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: the role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of personality and social psychology, 57*(6), 1082-1090.

Eagly, A. H., Chen, S., Chaiken, S., & Shaw-Barnes, K. (1999). The impact of attitudes on memory: an affair to remember. *Psychological Bulletin, 125*(1), 64-89.

Eagly, A. H., Kulesa, P., Brannon, L. A., Shaw, K., & Hutson-Comeaux, S. (2000). Why counterattitudinal messages are as memorable as proattitudinal messages: The importance of active defense against attack. *Personality and Social Psychology Bulletin, 26*(11), 1392.

Ebbesen, E. B., & Bowers, R. J. (1974). Proportion of risky to conservative arguments in a group discussion and choice shifts. *Journal of Personality and Social Psychology, 29*(3), 316-327.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71*, 5-24.

Evans, J. S. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology, 75*(4), 541-568.

Evans, J. S. B. T. (1989). *Bias in Human Reasoning: Causes and Consequences*. Hillsdale, NJ: Lawrence Erlbaum.

Evans, J. S. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87*, 223-240.

Evans, J. S. B. T. (1998). Matching Bias in Conditional Reasoning: Do We Understand it After 25 Years? *Thinking & Reasoning, 4*(1), 45-110.

Evans, J. S. B. T. (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychological bulletin, 128*(6), 978-996.

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13*(3), 378-395.

Evans, J. S. B. T. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgment*. Hove: Psychology Press.

Evans, J. S. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition, 11*, 295-306.

Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual process theory of reasoning. *Thinking and Reasoning, 11*, 382-389.

Evans, J. S. B. T., & Elqayam, S. (2007). Dual-processing explains base-rate neglect, but which dual-process theory and how? *Behavioral and Brain Sciences, 30*(03), 261-262.

Evans, J. S. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25*(6), 1495-1513.

Evans, J. S. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*(3), 391-397.

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Evans, J. S. B. T., & Over, D. E. (1996). *Rationality and Reasoning*. Hove: Psychology Press.

Evans, J. S. B. T., & Wason, P. C. (1976). Rationalization in a reasoning task. *British Journal of Psychology, 67*, 479-486.

Farnsworth, P. R., & Behner, A. (1931). A note on the attitude of social conformity. *Journal of Social Psychology, 2*, 126-128.

Finkbeiner, M., & Forster, K. I. (2008). Attention, intention and domain-specific processing. *Trends in Cognitive Science, 12*(2), 59-64.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83.

Fodor, J. (1983). *The Modularity of Mind*. Cambridge, Massachusetts: MIT Press.

Fodor, J. (2001). *The Mind Doesn't Work That Way*. Cambridge, Massachusetts: MIT Press.

Foot, H., Howe, C., Anderson, A., Tolmie, A., & Warden, D. (1994). *Group and Interactive Learning*. Southampton: Computational Mechanics Press.

Fox, F. V., & Staw, B. M. (1979). The trapped administrator: The effects of job insecurity and policy resistance upon commitment to a course of action. *Administrative Science Quarterly, 24*, 449-471.

Franklin, B. (1799). *The Autobiography of Benjamin Franklin*.

Frantz, C. M., & Janoff-Bulman, R. (2000). Considering both sides: The limits of perspective taking. *Basic and Applied Social Psychology, 22*, 31–42.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*(4), 25-42.

Frey, D., & Stahlberg, D. (1986). Selection of Information after Receiving more or Less Reliable Self-Threatening Information. *Personality and Social Psychology Bulletin, 12*(4), 434.

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., et al. (2007). Self-Control Relies on Glucose as a Limited Energy Source: Willpower Is More Than a Metaphor. *Journal of Personality and Social Psychology, 92*(2), 325.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692-731.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition, 86*(3), 223-251.

Gigerenzer, G. (2007). Fast and frugal heuristics: The tools of bounded rationality. In D. Koehler & N. Harvey (Eds.), *Handbook of Judgment and Decision Making*. Oxford, UK: Blackwell.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology, 59*(4), 601-613.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54*(5), 733-740.

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology, 65*(2), 221-233.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology, 44*(6), 1110-1126.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge, UK: Cambridge University Press.

Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of personality and social psychology, 52*(3), 464-474.

Glachan, M., & Light, P. (1982). Peer interaction and learning: Can two wrongs make a right? In G. Butterworth & P. Light (Eds.), *Social cognition: Studies in the development of understanding* (pp. 238–262). Chicago: University of Chicago Press.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Gouzoules, H., & Gouzoules, S. (2002). Primate communication: by nature honest, or by experience wise? *International Journal of Primatology, 23*(4), 821-848.

Gouzoules, H., Gouzoules, S., & Miller, K. (1996). Skeptical responding in rhesus monkeys (Macaca mulatta). *International Journal of Primatology, 17*, 549-568.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144â€“1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron, 44*(2), 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., Cohen, J. D., Mapping, B., et al. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105-2108.

Greenwald, A. G. (1969). The open-mindedness of the counterattitudinal role player. *Journal of Experimental Social Psychology, 5*(4), 375-388.

Greenwald, A. G., Spangenberg, E. R., Pratkanis, A. R., & Eskenazi, J. (1991). Double-Blind Tests of Subliminal Self-Help Audiotapes. *Psychological Science, 2*(2), 119-122.

Griggs, R. A., & Cox, J. R. (1983). The effects of problem content and negation on Wason's selection task. *Quarterly Journal of Experimental Psychology, 35A*, 519-533.

Gruter, C., Balbuena, M. S., & Farina, W. M. (2008). Informational conflicts created by the waggle dance. *Proceedings of the Royal Society B: Biological Sciences, 275*(1640), 1321-1327.

Guenther, C. L., & Alicke, M. D. (2008). Self-enhancement and belief perseverance. *Journal of Experimental Social Psychology, 44*(3), 706-712.

Guetzkow, H., & Gyr, J. (1954). An Analysis of Conflict in Decision-Making Groups. *Human Relations, 7*(3), 367.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological Review, 114*(3), 704-732.

Hahn, U., Oaksford, M., & Bayindir, H. (2005). How Convinced Should We Be by Negative Evidence? *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review, 108*(4), 8I4-834.

Halberstadt, J. B., & Levine, G. M. (1999). Effects of reasons analysis on the accuracy of predicting basketball games. *Journal of Applied Social Psychology, 29*(3), 517-530.

Hamilton, W. D. (1964a). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology, 7*(1), 1-16.

Hamilton, W. D. (1964b). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology, 7*(1), 17-52.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour, 61*, 139–151.

Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition, 101*(3), 495-514.

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skillful in competitive than in cooperative cognitive tasks. *Animal Behaviour, 68*, 571-581.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70*, 117-133.

Harvey, N., Harries, C., & Fischer, I. (2000). Using Advice and Assessing Its Quality. *Organizational Behavior and Human Decision Processes, 81*(2), 252-273.

Hasson, U., Simmons, J. P., & Todorov, A. (2005). Believe It or Not: On the Possibility of Suspending Belief. *Psychological Science, 16*(7), 566-571.

Hennessy, M. H., Fishbein, M., Curtis, B., & Barrett, D. (2008). Confirming preferences or collecting data? Information search strategies and romantic partner selection. *Psychology, Health & Medicine, 13*(2), 202.

Hill, G. W. (1982). Group versus individual performance: Are N + 1 heads better than one? *Psychological Bulletin, 91*, 517-539.

Hinsz, V. B., Tindale, R. S., & Nagao, D. H. (2008). Accentuation of information processes and biases in group judgments integrating base-rate and case-specific information. *Journal of Experimental Social Psychology, 44*(1), 116-126.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 719-731.

Holton, B., & Pyszczynski, T. (1989). Biased Information Search in the Interpersonal Domain. *Personality and Social Psychology Bulletin, 15*(1), 42.

Howe, C. J. (1990). Physics in the Primary School: Peer Interaction and the Understanding of Floating and Sinking. *European Journal of Psychology of Education, 5*(4), 459-475.

Hsee, C. K. (1995). Elastic Justification: How Tempting but Task-Irrelevant Factors Influence Decisions. *Organizational Behavior and Human Decision Processes, 62*(3), 330-337.

Hsee, C. K. (1996a). Elastic Justification: How Unjustifiable Factors Influence Judgments. *Organizational Behavior and Human Decision Processes, 66*(1), 122-129.

Hsee, C. K. (1996b). The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives. *Organizational Behavior and Human Decision Processes, 67*(3), 247-257.

Hsee, C. K. (1998). Less Is Better: When Low-value Options Are Valued More Highly than High-value Options. *Journal of Behavioral Decision Making, 11*.

Hsee, C. K. (1999). Value seeking and prediction-decision inconsistency: why don't people take what they predict they'll like the most? *Psychonomic Bulletin and Review, 6*(4), 555-561.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576-590.

Hsee, C. K., Yu, F., Zhang, J., & Zhang, Y. (2003). Medium Maximization. *Journal of Consumer Research, 30*(1), 1-14.

Hsee, C. K., & Zhang, J. (2004). Distinction Bias: Misprediction and Mischoice Due to Joint Evaluation. *Journal of Personality and Social Psychology, 86*(5).

Hsee, C. K., Zhang, J., Yu, F., & Xi, Y. (2003). Lay rationalism and inconsistency between predicted experience and decision. *Journal of Behavioral Decision Making, 16*(4), 257-272.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *The Journal of Consumer Research, 9*(1), 90-98.

Igou, E. R., & Bless, H. (2007). On undesirable consequences of thinking: framing effects as a function of substantive processing. *Journal of Behavioral Decision Making, 20*(2), 125.

Irwin, J. R., Slovic, P., Lichtenstein, S., & McClelland, G. H. (1993). Preference reversals and the measurement of environmental values. *Journal of Risk and Uncertainty, 6*(1), 5-18.

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology, 50*(6), 1141-1151.

James, W. (1950). *The Principles of Psychology*. New York: Dover.

Janis, I. L. (1982). *Groupthink* (2nd Rev. ed.). Boston: Houghton Mifflin.

Jehn, K. A. (1995). A Multimethod Examination of the Benefits and Detriments of Intragroup Conflict. *Administrative Science Quarterly, 40*(2).

Jehn, K. A., & Mannix, E. A. (2001). The dynamic nature of conflict: A longitudinal study of intragroup conflict and group performance. *Academy of Management Journal, 44*(2), 238-251.

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why Differences Make a Difference: A Field Study of Diversity, Conflict, and Performance in Workgroups. *Administrative Science Quarterly, 44*(4).

Jellison, J. M., & Mills, J. (1969). Effect of public commitment upon opinions. *Journal of Experimental Social Psychology, 5*(3), 340-346.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge, UK: Cambridge University Press.

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology, 50*, 109-135.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences, 5*, 434-442.

Johnson-Laird, P. N. (2006). *How We Reason*. Oxford: Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove: Lawrence Erlbaum Associates Ltd.

Johnson-Laird, P. N., & Byrne, R. M. J. (1996). Mental models and syllogisms. *Behavioural and Brain Sciences, 19*, 543-546.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109*, 646-678.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology, 10*, 64-99.

Johnson, B. T., & Eagly, A. H. (1989). Effects of involvement on persuasion: A meta-analysis. *Psychological Bulletin, 106*(2), 290-314.

Johnston, L. (1996). Resisting change: information-seeking and stereotype change. *European Journal of Social Psychology, 26*, 799-825.

Jonas, E., Greenberg, J., & Frey, D. (2003). Connecting Terror Management and Dissonance Theory: Evidence that Mortality Salience Increases the Preference for Supporting Information after Decisions. *Personality and Social Psychology Bulletin, 29*(9), 1181.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision, 50*(1), 59-99.

Joseph, J. S., Chun, M. M., & Nakayama, K. (1997). Attentional requirements in a 'preattentive' feature search task. *Nature 387*, 805-807.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 49-81). Cambridge, UK: Cambridge University Press.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–294). Cambridge, UK: Cambridge Univiversity Press.

Kahneman, D., & Ritov, I. (1994). Determinants of stated willingness to pay for public goods: A study in the headline method. *Journal of Risk and Uncertainty, 9*(1), 5-37.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430-454.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica, 47*(2), 263-292.

Kahneman, D., & Tversky, A. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Quarterly Journal of Economics, 106*(4), 1039-1061.

Kaplan, M. F., & Miller, C. E. (1977). Judgments and group discussion: Effect of presentation and memory factors on polarization. *Sociometry, 40*(4), 337-343.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*(4), 681-706.

Kerr, N. L., Maccoun, R. J., & Kramer, G. P. (1996). Bias in judgement: comparing individuals and groups. *Psychological review, 103*(4), 687-719.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*, 623-655.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kirby, K. N., & Herrnstein, R. J. (1995). Preference reversals due to myopic discounting of delayed reward. *Psychological Science, 6*(2), 83-89.

Kirschner, M., & Gerhart, J. (1998). Evolvability. *PNAS, 95*(15), 8420-8427.

Klaczynski, P. A. (1997). Bias in adolescents' everyday reasoning and its relationship with intellectual ability, personal theories, and self-serving motivation. *Developmental Psychology, 33*, 273-283.

Klaczynski, P. A., & Cottrell, J. M. (2004). A dual-process approach to cognitive development: The case of children's understanding of sunk cost decisions. *Thinking & Reasoning, 10*(2), 147-174.

Klaczynski, P. A., & Gordon, D. H. (1996a). Everyday Statistical Reasoning during Adolescence and Young Adulthood: Motivational, General Ability, and Developmental Influences. *Child Development, 67*(6), 2873-2891.

Klaczynski, P. A., & Gordon, D. H. (1996b). Self-serving influences on adolescents' evaluations of belief-relevant evidence. *Journal of Experimental Child Psychology, 62*, 317-339.

Klaczynski, P. A., Gordon, D. H., & Fauth, J. (1997). Goal-oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology, 89*, 470-485.

Klaczynski, P. A., & Lavallee, K. L. (2005). Domain-specific identity, epistemic regulation, and intellectual ability as predictors of belief-based reasoning: A dual-process perspective. *Journal of Experimental Child Psychology, 92*, 1-24.

Klaczynski, P. A., & Narasimham, G. (1998). Development of scientific reasoning biases: Cognitive versus ego-protective explanations. *Developmental Psychology, 34*, 175-187.

Klaczynski, P. A., & Robinson, B. (2000). Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning tasks. *Psychology and Aging, 15*, 400-416.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychol Rev, 107*(4), 852-884.

Klayman, J. (1995). Varieties of confirmation bias. In J. R. Busemeyer, R. Hastie & D. L. Medin (Eds.), *Decision Making from the Perspective of Cognitive Psychology* (pp. 385-418). New York: Academic Press.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Klein, G. (1998). *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press.

Koehler, J. J. (1993). The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality. *Organizational Behavior and Human Decision Processes, 56*, 28-28.

Kogan, N., & Wallach, M. A. (1966). Modification of a judgmental style through group interaction. *Journal of Personality and Social Psychology, 4*(2), 165-174.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory and Cognition, 6*, 107-118.

Kray, L. J., & Galinsky, A. D. (2003). The debiasing effect of counterfactual mind-sets: Increasing the search for disconfirmatory information in group decisions. *Organizational Behavior and Human Decision Processes, 91*(1), 69-81.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (2ème ed., pp. 390-402). Oxford: Basil Blackwell Scientific Publications.

Krueger, J. L. (2003). Return of the ego-self-referent information as a filter for social prediction: comment on Karniol (2003). *Psychological Review, 110*, 585-590.

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology, 19*(5), 448-468.

Kuhn, D. (1991). *The Skills of Arguments*. Cambridge: Cambridge University Press.

Kuhn, D., & Lao, J. (1996). Effects of Evidence on Attitudes: Is Polarization the Norm? *Psychological Science, 7*, 115-120.

Kuhn, D., Shaw, V. F., & Felton, M. (1997). Effects of dyadic interaction on argumentative reasoning. *Cognition and Instruction, 15*, 287-315.

Kuhn, D., Weinstock, M., & Flaton, R. (1994). How well do jurors reason? Competence

dimensions of individual variation in a juror reasoning task. *Psychological Science, 5*, 289–296.

Kunda, Z. (1987). Motivation and inference: Self-serving generation and evaluation of evidence. *Journalof Personality and Social Psychology, 53*(636-647).

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498.

Kunda, Z., Fong, G. T., Sanitioso, R., & Reber, E. (1993). Directional questions direct self-conceptions. *Journal of Experimental Social Psychology, 29*(1), 63-86.

Kurzban, R., & Aktipis, A. (2007). Modularity and the Social Mind: Are Psychologists Too Self-ish? *Personality and Social Psychology Review, 11*(2), 131.

Lambert, A. J., Cronen, S., Chasteen, A. L., & Lickel, B. (1996). Private vs public expressions of racial prejudice. *Journal of Experimental Social Psychology, 32*(5), 437-459.

Latane, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37*(6), 822-832.

Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes, 88*, 605-620.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22*, 177–189.

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology, 90*, 644–651.

Laughlin, P. R., Kerr, N. L., Davis, J. H., Halff, H. M., & Marciniak, K. A. (1975). Group size, member ability, and social decision schemes on an intellective task. *Journal of Personality and Social Psychology, 33*, 80-88.

Laughlin, P. R., VanderStoep, S. W., & Hollingshead, A. B. (1991). Collective versus individual induction: Recognition of truth, rejection of error, and collective information processing. *Journal of Personality and Social Psychology, 61*, 50-67.

Laughlin, P. R., Zander, M. L., Knievel, E. M., & Tan, T. S. (2003). Groups perform better than the best individuals on letters-to-numbers problems: Informative equations and effective reasoning. *Journal of Personality and Social Psychology, 85*, 684-694.

LeBoeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: on the susceptibility to framing effects. *Journal of Behavioral Decision Making, 16*(2), 77-92.

Leising, K. J., Wong, J., Waldmann, M. R., & Blaisdell, A. P. (2008). The special status of actions in causal reasoning in rats. *Journal of Experimental Psychology: General, 137*(3), 514-527.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255-275.

Levine, G. M., Halberstadt, J. B., & Goldstone, R. L. (1996). Reasoning and the Weighting of Attributes in Attitude Judgments. *Journal of Personality and Social Psychology, 70*, 230-240.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9-16.

Liberman, A., & Chaiken, S. (1991). Value conflict and thought-induced attitude change. *Journal of Experimental Social Psychology, 27*(3), 203-216.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334306-334334). New York: Cambridge University Press.

Lien, M. C., Ruthruff, E., Cornett, L., Goodin, Z., & Allen, P. A. (2008). On the nonautomaticity of visual word processing: electrophysiological evidence that word processing requires central attention. *Journal of Experimental Psychology: Human Perception and Performance, 34*(3), 751-773.

Littlepage, G. E., & Mueller, A. L. (1997). Recognition and utilization of expertise in problem-solving groups: Expert characteristics and behavior. *Group Dynamics, 1*, 324-328.

Loewenstein, G., & Prelec, D. (1998). The red and the black: Mental accounting of savings and debt. *Marketing Science, 17*(4), 4-28.

Lombardelli, C., Proudman, J., & Talbot, J. (2005). Committees versus individuals: An experimental analysis of monetary policy decision-making. *International Journal of Central Banking, May*, 181-205.

Lopez, D. F., Ditto, P. H., & Waghorn, K. C. (1994). Valenced social information and the temporal location of thought. *British Journal of Social Psychology Quarterly, 33*, 443-456.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*, 1231-1243.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098-2109.

Lucas, E. J., & Ball, L. J. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking and Reasoning, 11*, 35-66.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology, 15*(3), 215-233.

Maciejovsky, B., & Budescu, D. V. (2007). Collective induction without cooperation? Learning and knowledge transfer in cooperative groups and competitive auctions. *Journal of personality and social psychology, 92*(5), 854-870.

Madsen, D. B. (1978). Issue importance and group choice shifts: A persuasive arguments approach. *Journal of Personality and Social Psychology, 36*, 1118-1127.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161-175.

Maloney, L. T., Trommershauser, J., & Landy, M. S. (2007). Questions without words: A comparison between decision making under risk and movement planning under risk. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 297â€“313). New York: Oxford University Press.

Markman, E. M., & Gorin, L. (1981). Children's ability to adjust their standards for evaluating comprehension. *Journal of Educational Psychology, 73*(3), 320-325.

Markus, H., & Kunda, Z. (1986). Stability and malleability of the self-concept. *Journal of Personality and Social Psychology, 51*(4), 858-866.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman.

Masicampo, E. J., & Baumeister, R. F. (2008). Toward a Physiology of Dual-Process Reasoning and Judgment: Lemonade, Willpower, and Expensive Rule-Based Analysis. *Psychological Science, 19*(3), 255-260.

Matthew, R., & Schrag, J. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics, 114*, 37-82.

Maynard Smith, J., & Harper, D. (2003). *Animal Signals*. Oxford: Oxford University Press.

Mazar, N., Amir, O., & Ariely, D. (In prep). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance.

McCoy, M. L., Nunez, N., & Dammeyer, M. M. (1999). The Effect of Jury Deliberations on Jurors' Reasoning Skills. *Law and Human Behavior, 23*(5), 557-575.

McGuire, T. W., Kiesler, S., & Siegel, J. (1987). Group and computer-mediated discussion effects in risk decision making. *Journal of Personality and Social Psychology, 52*(5), 917-930.

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1). New York: Academic Press.

McMackin, J., & Slovic, P. (2000). When does explicit justification impair decision making? *Journal of Applied Cognitive Psychology, 14*, 527–541.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10*(3), 517-532.

Mercier, H., & Sperber, D. (in press). Intuitive and reflective inferences. In J. S. B. T. Evans & K. Frankish (Eds.), *In Two Minds*. New York: Oxford University Press.

Mercier, H., Van der Henst, J.-B., Yama, H., Kawasaki, Y., & Adachi, K. (submitted). Strategies for taking advice into account: a cross-cultural study.

Michaelsen, L. K., Watson, W. E., & Black, R. H. (1989). A realistic test of individual versus group consensus decision making. *Journal of Applied Psychology, 74*(5), 834-839.

Milgram, S. (1974). *Obedience to Authority: An Experimental View*. New York: Harper & Row.

Millar, M. G., & Tesser, A. (1986). Thought-induced attitude change: the effects of schema structure and commitment. *Journal of Personality and Social Psychology, 51*(2), 259-269.

Millar, M. G., & Tesser, A. (1989). The effects of affective-cognitive consistency and thought on the attitude-behavior relation. *Journal of Experimental Social Psychology, 25*, 189-202.

Miller, A. G., Michoskey, J. W., Bane, C. M., & Dowd, T. G. (1993). The attitude polarization phenomenon: role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology, 64*(4), 561-574.

Molden, D. C., & Higgins, E. T. (2008). How preferences for eager versus vigilant judgment strategies affect self-serving conclusions. *Journal of Experimental Social Psychology, 44*(5), 1219-1228.

Montesquieu. (1784). *De l'Esprit des lois*

Moore, T. E. (1982). Subliminal advertising: What you see is what you get. *Journal of Marketing, 46*(2), 38-47.

Morsanyi, K., & Handley, S. J. (2008). How smart do you need to be to get it wrong? The role of cognitive capacity in the development of heuristic-based judgment. *Journal of Experimental Child Psychology, 99*(1), 18-36.

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning, 4*(3), 231-248.

Mullainathan, S., & Shleifer, A. (2005). Persuasion in finance. *NBER Working Paper*.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule Learning by Rats. *Science, 319*(5871), 1849.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The Quarterly Journal of Experimental Psychology, 29*(1), 85-95.

Navarro, A. D., & Fantino, E. (2005). The Sunk Cost Effect In Pigeons And Humans. *Journal of the Experimental Analysis of Behavior, 83*(1), 1.

Nemeth, C., & Rogers, J. (1996). Dissent and the search for information. *British Journal of Social Psychology, 35*, 67-76.

Neuman, Y. (2003). Go ahead, prove that God does not exist! On high school students' ability to deal with fallacious arguments. *Learning and Instruction, 13*(4), 367-380.

Neuman, Y., Glassner, A., & Weinstock, M. (2004). The effect of a reason's truth-value on the judgment of a fallacious argument. *Acta Psychologica, 116*(2), 173-184.

Neuman, Y., Weinstock, M. P., & Glasner, A. (2006). The effect of contextual factors on the judgement of informal reasoning fallacies. *The Quarterly Journal of Experimental Psychology, 59*(2), 411-425.

Newell, B. R., Wong, K. Y., Cheung, J. C. H., & Rakow, T. (In Press). Think, blink or sleep on it? The impact of modes of thought on complex decision making. *The Quarterly Journal of Experimental Psychology*.

Newstead, S. E. (1989). Interpretational errors in syllogistic reasoning. *Journal of Memory and Language, 28*, 78-91.

Newstead, S. E. (1995). Gricean implicatures and syllogistic reasoning. *Journal of Memory and Language, 34*(644-664).

Newstead, S. E., Handley, S. J., & Buck, E. (1999). Falsifying mental models: testing the predictions of theories of syllogistic reasoning. *Memory and Cognition, 27*(2), 344-354.

Nisbett, R. E., & Wilson, T. (1977). Telling more than we can know. *Psychological Review, 84*(1), 231-259.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences, 12*(7), 265-272.

Noveck, I., Van der Henst, J.-B., Rossi, S., & Mercier, H. (2007). Psychologie cognitive et raisonnement. In S. Rossi & J.-B. Van der Henst (Eds.), *Psychologies du raisonnement*. Bruxelles: DeBoeck.

Nozick, R. (1993). *The Nature of Rationality*. New York: Princeton University Press.

Nussbaum, E. M., & Sinatra, G. M. (2003). Argument and conceptual engagement. *Contemporary Educational Psychology, 28*(3), 384-395.

Nyhan, B., & Reifler, J. (In prep.). When Corrections Fail.

Oaksford, M., Chater, N., & Grainger, R. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5*, 193-243.

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology, 58*(2), 75-85.

Okada, E. M. (2005). Justification effects on consumer choice of hedonic and utilitarian goods. *Journal of Marketing Research, 42*(1), 43-53.

Okada, E. M., & Hoch, S. J. (2004). Spending time versus spending money. *Journal of Consumer Research, 31*(2), 313-323.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*.

Otero, J., & Kintsch, W. (1992). Failures to detect contradiction in a text: What readers believe versus what they read. *Psychological Science, 3*(4), 229-235.

Packard, V. (1957). *The Hidden Persuaders*. New York: David McKay.

Paese, P. W., Bieser, M., & Tubbs, M. E. (1993). Framing Effects and Choice Shifts in Group Decision Making. *Organizational Behavior and Human Decision Processes, 56*, 149-149.

Pascal, B. (1670). *Pensées*.

Pelham, B. W., & Neter, E. (1995). The effect of motivation of jugdment depends on the difficulty of the judgment. *Journal of personality and social psychology, 68*(4), 581-594.

Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision-making. *Cognition, 49*, 123-163.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology, 77*, 562-571.

Perkins, D. N. (1989). *Reasoning as it is and could be: An empirical perspective.* Paper presented at the Thinking across cultures: The third international conference on thinking.

Pessoa, L., Kastner, S., & Ungerleider, L. G. (2002). Attentional control of the processing of neural and emotional stimuli. *Brain Research Cognitive Brain Research, 15*(1), 31-45.

Pessoa, L., McKenna, M., Gutierrez, E., & Ungerleider, L. G. (2002). Neural processing of emotional faces requires attention. *PNAS, 99*, 11458-11463.

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology, 37*, 349-360.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 19, pp. 123-205). Orlando, FL: Academic Press.

Petty, R. E., & Cacioppo, J. T. (1990). Involvement and persuasion: Tradition versus integration. *Psychological Bulletin, 107*(3), 367-374.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41*(5), 847-855.

Petty, R. E., & Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. In D. Gilbert, S. Fiske & G. Lindzey (Eds.), *The Handbook of Social Psychology* (Vol. 1, pp. 323–390). Boston: McGraw-Hill.

Pigliucci, M. (2008). Is evolvability evolvable? *Nature Review Genetics, 9*(1), 75-82.

Pinker, S. (2000). *Comment fonctionne l'esprit*. Paris: Odile Jacob.

Pinkley, R. L., Griffith, T. L., & Northcraft, G. B. (1995). " Fixed Pie" a la Mode: Information Availability, Information Processing, and the Negotiation of Suboptimal Agreements. *Organizational Behavior and Human Decision Processes, 62*(1), 101-112.

Poletiek, F. H. (1996). Paradoxes of falsification. *Quarterly Journal of Experimental Psychology, 49A*, 447-462.

Politzer, G., & Mercier, H. (In press). Solving categorical syllogisms with singular premises. *Thinking and Reasoning*.

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. (1995). Attitude strength and resistance processes. *Journal of Personality and Social Psychology, 69*(3), 408-419.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium*. Hillsdale, NJ: Erlbaum.

Pratkanis, A. R., & Aronson, E. (1992). *Age of Propaganda: The Everyday Use and Abuse of Persuasion*. New York: W.H. Freeman and Company.

Prescott, T. J., Redgrave, P., & Gurney, K. N. (1999). Layered control architectures in robots and vertebrates. *Adaptive Behavior, 7*(1), 99-127.

Pyszczynski, T., & Greenberg, J. (1987). Toward and integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 20, pp. 297-340). New York: Academic Press.

Ratneshwar, S., Shocker, A. D., & Stewart, D. W. (1987). Toward Understanding the Attraction Effect: The Implications of Product Stimulus Meaningfulness and Familiarity. *Journal of Consumer Research, 13*(4), 520.

Real, L. A. (1992). Information processing and the evolutionary ecology of cognitive architecture. *The American Naturalist, 140*, S108-S145.

Reber, A. S. (1993). *Implicit Learning and Tacit Knowledge*. New York: Oxford University Press.

Redgrave, P., Prescott, T. J., & Gurney, K. N. (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience, 89*, 1009-1023.

Redlawsk, D. P. (2002). Hot Cognition or Cool Consideration? Testing the Effects of Motivated Reasoning on Political Decision Making. *The Journal of Politics, 64*(4), 1021-1044.

Rees, G., Frith, C. D., & Lavie, N. (1997). Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science 278* 1616-1619.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99): Appleton-Century-Crofts.

Resnick, L. B., Salmon, M., Zeitz, C. M., Wathen, S. H., & Holowchak, M. (1993). Reasoning in conversation. *Cognition and Instruction, 11*(3/4), 347-364.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*(1), 1-75.

Ricco, R. B. (2003). The macrostructure of informal arguments: A proposed model and analysis. *The Quarterly Journal of Experimental Psychology A, 56*(6), 1021-1051.

Rips, L. J. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA: MIT Press.

Rips, L. J. (1998). Reasoning and conversation. *Psychological Review, 105*, 411–441.

Rips, L. J. (2002). Circular reasoning. *Cognitive Science, 26*, 767–795.

Roberts, M. J., & Newton, E. J. (2002). Inspection times, the change task, and the rapid response selection task. *Quarterly Journal of Experimental Psychology, 54*, 1031-1048.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm. *Journal of Personality and Social Psychology, 32*(5), 880-802.

Ross, M., McFarland, C., & Fletcher, G. J. (1981). The effect of attitude on the recall of personal histories. *Journal of Personality and Social Psychology, 40*(4), 627â€“634.

Sadler, O., & Tesser, A. (1973). Some effects of salience and time upon interpersonal hostility and attraction during social isolation. *Sociometry, 36*(1), 99-112.

Sanitioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social Psychology, 59*(2), 229-241.

Santangelo, V., Belardinelli, O., & Spence, C. (2007). The suppression of reflexive visual and auditory orienting when attention is otherwise engaged. *Journal of Experimental Psychology: Human Perception and Performance, 33*(1), 12.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Schacter, D. L. (1987). Implicit Memory: History and Current Status. *Journal of experimental psychology. Learning, memory, and cognition, 13*(3), 501-518.

Schkade, D., Sunstein, C. R., & Kahneman, D. (2000). Deliberating about dollars: The severity shift. *Columbia Law Review, 100*, 1139–1176.

Schooler, J. W., & Engster-Schooler, T. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36-71.

Schroyens, W., Schaeken, W., & Handley, S. (2003). In search of counter-examples: Deductive rationality in human reasoning. *The Quarterly Journal of Experimental Psychology Section A, 56*(7), 1129-1145.

Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., & Frey, D. (2006). Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *Journal of Personality and Social Psychology, 91*(6), 1080-1093.

Schwartz, B. (2004). *The Paradox of Choice: Why More is Less*. New York: Harper Collins.

Schwarz, B. B., Neuman, Y., & Biezuner, S. (2000). Two wrongs make a right. . .if they argue together! *Cognition and Instruction, 18*, 461–494.

Schwarz, N. (1991). Feelings as information: Informational and motivational functions of affective states. In E. T. Higgins & R. M. Sorrentino (Eds.), *The Handbook of Motivation and Cognition: Foundations of Social Behavior* (Vol. 2, pp. 527-561). New York: Guilford Press.

Schwarz, N. (2004). Metacognitive Experiences in Consumer Judgment and Decision Making. *Journal of Consumer Psychology, 14*(4), 332-348.

Schweitzer, M. E., & Hsee, C. K. (2002). Stretching the Truth: Elastic Justification and Motivated Communication of Uncertain Information. *Journal of Risk and Uncertainty, 25*(2), 185-201.

Sears, D. O., Freedman, J. L., & O'Connor, E. F. (1964). The Effects of Anticipated Debate and Commitment on the Polarization of Audience Opinion. *Public Opinion Quarterly, 28*(4), 615-627.

Selart, M. (1996). Structure Compatibility and Restructuring in Judgment and Choice. *Organizational Behavior and Human Decision Processes, 65*(2), 106-116.

Sengupta, J., & Fitzsimons, G. J. (2004). The effect of analyzing reasons on the stability of brand attitudes: A reconciliation of opposing predictions. *Journal of Consumer Research, 31*(3), 705-711.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition, 21*, 546-546.

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition, 49*(1-2), 11-36.

Shafir, E., & Tversky, A. (1992). Thinking through Uncertainty: Nonconsequential Reasoning and Choice. *Cognitive Psychology, 24*(4), 449-474.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27*, 379-423, 623-656.

Shaw, V. F. (1996). The Cognitive Processes in Informal Reasoning. *Thinking & Reasoning, 2*(1), 51-80.

Sherman, P. W. (1977). Nepotism and the Evolution of Alarm Calls. *Science, 197*(4310), 1246-1253.

Silk, J. B., Kaldor, E., & Boyd, R. (2000). Cheap talk when interests conflict. *Animal Behavior, 59*, 423-432.

Simons, T. L., & Peterson, R. S. (2000). Task conflict and relationship conflict in top management teams: The pivotal role of intragroup trust. *Journal of Applied Psychology, 85*(1), 102-111.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *The Journal of Consumer Research, 16*(2), 158-174.

Simonson, I., Carmon, Z., & O'Curry, S. (1994). Experimental evidence on the negative effect of product features and sales promotions on brand choice. *Marketing Science, 13*, 23-23.

Simonson, I., & Nowlis, S. M. (2000). The role of explanations and need for uniqueness in consumer decision making: Unconventional choices based on reasons. *Journal of Consumer Research, 27*(1), 49-68.

Simonson, I., Nowlis, S. M., & Simonson, Y. (1993). The Effect of Irrelevant Preference Arguments on Consumer Choice. *Journal of Consumer Psychology, 2*(3), 287-306.

Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes, 51*(3), 416-446.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25*(2), 231-280.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3-22.

Slovic, P. (1975). Choice between equally valued alternatives. *Journal of Experimental Psychology: Human Perception and Performance, 1*(3), 280-287.

Smith, A. (1978). *Lectures on Jurisprudence*. Oxford: Oxford University Press.

Smith, S. M., Fabrigar, L. R., & Norris, M. E. (2008). Reflecting on Six Decades of Selective Exposure Research: Progress, Challenges, and Opportunities. *Social and Personality Psychology Compass, 2*(1), 464-493.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational behavior and human decision processes(Print), 43*(1), 1-28.

Snyder, C. R., & Fromkin, H. L. (1977). Abnormality as a Positive Characteristic: The Development and Validation of a Scale Measuring Need for Uniqueness. *Journal of Abnormal Psychology, 86*(5), 518-527.

Snyder, M., & Cantor, N. (1979). Testing hypotheses about other people: The use of historical knowledge. *Journal of Experimental Social Psychology, 15*(2), 330-342.

Snyder, M., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: an attributional ambiguity analysis. *Journal of Personality and Social Psychology, 37*(12), 2297-2306.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36*(11), 1202-1212.

Soman, D., & Cheema, A. (2001). The Effect of Windfall Gains on the Sunk-Cost Effect. *Marketing Letters, 12*(1), 51-62.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 39-67). Cambridge: Cambridge University Press.

Sperber, D. (1997). Intuitive and reflective beliefs. *Mind and Language, 12*(1), 67-83.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.), *Metarepresentations: A Multidisciplinary Perspective* (pp. 117-137). Oxford: Oxford University Press.

Sperber, D. (2001). An evolutionary perspective on testimony and argumentation. *Philosophical Topics, 29*, 401-413.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57*, 31-95.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (in prep). Epistemic vigilance.

Sperber, D., & Mercier, H. (in prep). Reasoning as a social activity. In J. Elster (Ed.), *Collective wisdom*.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Spinoza, B. (1677). *Ethics*.

Stanovich, K. E. (1999). *Who Is Rational?: Studies of Individual Differences in Reasoning*. Mahwah, NJ: Lawrence Erlbaum.

Stanovich, K. E. (2004). *The Robot's Rebellion*. Chicago: Chicago University Press.

Stanovich, K. E., & West, R. F. (1999). Discrepancies Between Normative and Descriptive Models of Decision Making and the Understanding/Acceptance Principle. *Cognitive Psychology, 38*(3), 349-385.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645-726.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning, 13*(3), 225-247.

Stanovich, K. E., & West, R. F. (2008a). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*(4), 672-695.

Stanovich, K. E. & West, R. F. (2008b). On the failure of cognitive ability to predict myside and one-sided thinking biases. *Thinking & Reasoning, 14*(2), 129-167.

Stasson, M. F., Kameda, T., Parks, C. D., Zimmerman, S. K., & Davis, J. H. (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly, 54*, 25-35.

Stayman, D. M., & Kardes, F. R. (1992). Spontaneous Inference Processes in Advertising. *Journal of Consumer Psychology, 1*(2), 125-142.

Stein, N. L., & Albro, E. R. (2001). The origins and nature of arguments: Studies in conflict understanding, emotion, and negotiation. *Discourse Processes, 32*(2&3), 113-133.

Stein, N. L., & Bernas, R. (1999). The early emergence of argumentative knowledge and skill. In J. Andriessen & P. Coirier (Eds.), *Foundations of Argumentative Text Processing* (pp. 97-116). Amsterdam: Amsterdam University Press.

Stein, N. L., Bernas, R. S., & Calicchia, D. J. (1997). Conflict talk: Understanding and resolving arguments. In T. Givon (Ed.), *Conversation: Cognitive, Communicative and Social Perspectives*. Amsterdam: John Benjamins.

Stein, N. L., Bernas, R. S., Calicchia, D. J., & Wright, A. (1995). Understanding and resolving arguments: The dynamics of negotiation. In B. Britton & A. G. Graesser (Eds.), *Models of Understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Stein, N. L., & Miller, C. A. (1993). The development of meaning and reasoning skill in argumentative contexts: Evaluating, explaining, and generating

evidence. In R. Glaser (Ed.), *Advances in Instructional Psychology* (Vol. 4, pp. 285-335). Hillsdale: Lawrence Erlbaum Associates.

Steiner, I. D. (1972). *Group processes and productivity*: New York: Academic Press.

Störring, G. (1908). Experimentelle untersuchungen über einfache Schlussprozesse. *Archiv für die Gesamte Psychologie, 11*, 1-127.

Strahan, E. J., Spencer, S. J., & Zanna, M. P. (2002). Subliminal priming and persuasion: Striking while the iron is hot. *Journal of Experimental Social Psychology, 38*(6), 556-568.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy, 10*(2), 175-195.

Sunstein, C. R. (2003). *Why Societies Need Dissent*. Cambridge: Harvard University Press.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50*(3), 755-769.

Tesser, A. (1976). Attitude Polarization as a Function of Thought and Reality Constraints. *Journal of Research in Personality, 10*(2), 183-194.

Tesser, A. (1978). Self-generated attitude change. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 11, pp. 289-338). New York: Academic Press.

Tesser, A., & Conlee, M. C. (1975). Some effects of time and thought on attitude polarization. *Journal of Personality and Social Psychology, 31*(2), 262-270.

Tesser, A., & Leone, C. (1977). Cognitive Schemas and Thought as Determinants of Attitude Change. *Journal of Experimental Social Psychology, 13*(4), 340-356.

Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology, 57*(3), 388-398.

Tetlock, P. E., Lerner, J. S., & Boettger, R. (1996). The dilution effect: Judgmental bias, conversational convention, or a bit of both? *European Journal of Social Psychology, 26*(6), 915-934.

Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: conformity, complexity, and bolstering. *Journal of Personality and Social Psychology, 57*(4), 632-640.

Thaler, R. (1980). Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior and Organization, 1*(1), 39-60.

Thaler, R., & Sunstein, C. (2007). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press.

Thompson, V. A. (1996). Reasoning from false premises The role of soundness in making logical deductions. *Canadian Journal of Experimental Psychology, 50*(3), 315-319.

Thompson, V. A., Evans, J. S. B. T., & Handley, S. J. (2005). Persuading and dissuading by conditional argument. *Journal of Memory and Language, 53*(2), 238-257.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., & Campbell, J. I. (2005). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin and Review, 10*(1), 184-189.

Tinbergen, N. (1963). On aims and methods in ethology. *Zeitschrift für Tierpsychologie, 20*, 410-433.

Tindale, R. S., & Sheffey, S. (2002). Shared information, cognitive load, and group memory. *Group Processes & Intergroup Relations, 5*(1), 5.

Tolmie, A., Howe, C., Mackenzie, M., & Greer, K. (1993). Task design as an influence on dialogue and learning: primary school group work with object flotation. *Social Development, 2*(3), 183-201.

Tomasello, M., Call, J., Nagell, C., Olguin, R., & Carpenter, M. (1994). The learning and use of gestural signals by young chimpanzees: A trans-generational study. *Primates, 35*, 137-154.

Tomasello, M., Gust, D., & Frost, T. A. (1989). A longitudinal investigation of gestural communication in young chimpanzees. *Primates, 30*, 35-50.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides & J. Tooby (Eds.), *The Adapted Mind* (pp. 19-136). Oxford: Oxford University Press.

Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology, 17*, 851-860.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*(1), 35-57.

Trognon, A. (1993). How does the process of interaction work when two interlocutors try to resolve a logical problem? *Cognition and Instruction, 11*(3&4), 325-345.

Trommershauser, J., Landy, M. S., & Maloney, L. T. (2006). Humans Rapidly Estimate Expected Gain in Movement Planning. *Psychological Science, 17*(11), 981-988.

Trommershauser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences, 12*(8), 291-297.

Tversky, A., & Griffin, D. (1991). Endowment and contrast in judgments of well-being. In F. Strack, M. Argyle & N. Schwarz (Eds.), *Subjective well-being: An interdisciplinary perspective* (Vol. 21, pp. 101-118). Oxford: Pergamon Press.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293-315.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95*(3), 371-384.

Tversky, A., & Shafir, E. (1992a). Choice under conflict: the dynamics of deferred decision. *Psychological Science, 3*(6), 358-361.

Tversky, A., & Shafir, E. (1992b). The disjunction effect in choice under uncertainty. *Psychological Science, 3*(5), 305-309.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science, 39*(10), 1179-1189.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., et al. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology, 32*(1), 109-123.

Uleman, J. (1999). Spontaneous versus intentional inferences in impression formation. In S. Chaiken & Y. Trope (Eds.), *Dual-Process Theories in Social Psychology*. New York: The Guilford Press.

Valdesolo, P., & DeSteno, D. (2007). Moral Hypocrisy: Social Groups and the Flexibility of Virtue. *Psychological Science, 18*(8), 689-690.

Webley, P., & Plaisier, Z. (1997). *Mental accounting in childhood*. Paper presented at the 16th Bi-Annual Conference on Subjective Probability, Utility, and Decision Making.

Weinstock, M., Neuman, Y., & Tabak, I. (2004). Missing the point or missing the norms? Epistemological norms as predictors of students' ability to identify fallacious arguments. *Contemporary Educational Psychology, 29*(1), 77-94.

Weir, W. (1984). Another look at subliminal "facts". *Advertising Age, October, 15*, 46.

Whiten, A., & Byrne, R. W. (Eds.). (1997). *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge: Cambridge University Press.

Whittlesea, B. W. A., & Williams, L. D. (2001a). The Discrepancy-Attribution Hypothesis: I. The Heuristic Basis of Feelings of Familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 3-13.

Whittlesea, B. W. A., & Williams, L. D. (2001b). The discrepancy-attribution hypothesis: II. Expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 14â€“33.

Whyte, G. (1993). Escalating commitment in individual and group decision making: A prospect theory approach. *Organizational Behavior and Human Decision Processes, 54*(3), 430-455.

Williams, G. C. (1966). *Adaptation and Natural Selection*. Princeton: Princeton University Press.

Wilson, T. D., Dunn, D. S., Bybee, J. A., Hyman, D. B., & Rotondo, J. A. (1984). Effects of analyzing reasons on attitude-behavior consistency. *Journal of Personality and Social Psychology, 47*(1), 5-16.

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 19, pp. 123-205). Orlando, FL: Academic Press.

Wilson, T. D., Kraft, D., & Dunn, D. S. (1989). The disruptive effects of explaining attitudes: the moderating effect of knowledge about the attitude object. *Journal of Experimental Social Psychology, 25*(5), 379-400.

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology*.

van Boxtel, C., van der Linden, J., & Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. *Learning and Instruction, 10*(4), 311-330.

Van der Henst, J.-B., Mercier, H., Yama, H., Kawasaki, Y., & Adachi, K. (2007). Dealing with contradiction in a communicative context: A cross-cultural study. *Intercultural Pragmatics*.

Van der Henst, J. -B. (2006). Relevance effects in reasoning. *Mind & Society, 5*(2), 229-245.

Vauvenargues. (1746). *Réflexions et maximes*

Vinokur, A. (1971). Review and theoretical analysis of the effects of group processes upon individual and group decisions involving risk. *Psychological Bulletin, 76*(4), 231-250.

Vinokur, A., & Burnstein, E. (1978). Depolarization of attitudes in groups. *Journal of Personality and Social Psychology, 36*(8), 872-885.

Vosniadou, S., Pearson, P. D., & Rogers, T. (1988). What causes children's failures to detect inconsistencies in test? Representation versus comparison difficulties. *Journal of Educational Psychology, 80*(1), 27-39.

Wagner, G. P., & Altenberg, L. (1996). Complex adaptations and the evolution of evolvability. *Evolution, 50*(3), 967-976.

Walster, E., Berscheid, E., Abrahams, D., & Aronson, V. (1967). Effectiveness of debriefing following deception experiments. *Journal of Personality and Social Psychology, 6*(4), 371-380.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-137.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology: I* (pp. 106–137). Harmandsworth, England: Penguin.

Wason, P. C. (1969). Structural simplicity and psychological complexity: some thoughts on a novel problem. *Bulletin of the British Psychological Society, 22*(28), 284.

Wason, P. C., & Evans, J. S. B. T. (1975). Dual processes in reasoning? *Cognition, 3*, 141-154.

Wilson, T. D., & LaFleur, S. J. (1995). Knowing what you'll do: effects of analyzing reasons on self-prediction. *J Pers Soc Psychol, 68*(1), 21-35.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101-126.

Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J., & LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin, 19*(3), 331.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Thinking, 60*(2), 181-192.

Wolfe, C. R. (2007). The locus of the myside bias in written argumentation. *Thinking & Reasoning, 14*(1), 1-27.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*(7-8), 1317-1329.

Wyer, R. S. J. R., & Frey, D. (1983). The effects of feedback about self and others on the recall and sudgments of feedback-relevant information. *Journal of Experimental Social Psychology, 19*(6), 540-559.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*, 1-13.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83*, 260-281.

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.