Le raisonnement comme compétence sociale : une comparaison expérimentale avec les théories intellectualistes

Emmanuel Trouche

N°d'ordre NNT : 2016 LYSE1132

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale N° 476
Neuroscience et Cognition

Spécialité de doctorat : SCIENCES COGNITIVES

Soutenue publiquement le 15/09/2016, par :

Emmanuel Trouche

Le raisonnement comme compétence sociale. Une comparaison expérimentale avec les théories intellectualistes.

Devant le jury composé de :

Sander, Emmanuel	Professeur - Université Paris 8	Président du jury
Bonnefon, Jean-François	Directeur de Recherche - CNRS	Rapporteur
De Neys, Wim	Chargé de Recherche - CNRS	Rapporteur
Fourneret, Pierre	Professeur - Université Lyon 1	Examinateur
Bazalgette, Didier	Direction Générale de l'Armement	Examinateur
Mercier, Hugo	Chargé de Recherche CNRS	Directeur de thèse

RESUME

La majorité des recherché, en Sciences Cognitives, suppose que la fonction du raisonnement humain est d'aider l'individu à avoir de meilleures croyances et à prendre de meilleures décisions, en particulier grâce à des mécanismes mentaux d'inférences logiques. En 2011, Dan Sperber et Hugo Mercier ont proposé une vision alternative du raisonnement humain. La fonction du raisonnement serait argumentative : le raisonnement serait ce qui permet aux individus de produire et d'évaluer des arguments en contextes dialogiques. Cette thèse a, d'une part, pour objectif de proposer une comparaison théorique entre les théories standards du raisonnement et la théorie argumentative du raisonnement. D'autre part, elle apporte un soutien empirique a la théorie argumentative à travers différents paradigmes expérimentaux (i.e., résolution de problème individuelle, production et évaluation d'arguments individuelle, résolution de problème et échange d'arguments en groupe). Cette thèse défend non seulement la valeur explicative de la théorie argumentative du raisonnement, mais caractérise également les mécanismes cognitifs du raisonnement humain, de par leurs fonctions, leurs biais, et les contextes qui les déclenchent.

MOTS CLÉS : Raisonnement – Argumentation – Biais Cognitifs – Résolution de problème

---

ABSTRACT

Most research in cognitive science assumes that the function of human reasoning is to help individual to improve their beliefs and make better decisions, in particular through mental mechanisms of logical inference. In 2011, Dan Sperber and Hugo Mercier put forward an alternative view of human reasoning. The function of reasoning would be argumentative: reasoning would be what enables individuals to produce and evaluate arguments in dialogical contexts. This PhD thesis aims at proposing a theoretical comparison between standard theories of reasoning and the argumentative theory of reasoning. Furthermore, it provides empirical support for the latter by using different experimental paradigms (i.e., individual problem solving, production and evaluation of arguments in solitary contexts, problem solving and arguments exchange in group). This thesis not only defends the explanatory value of the argumentative theory but also characterizes the cognitive mechanisms of human reasoning by their functions, their biases, and their triggering contexts.

KEYWORDS: Reasoning – Argumentation – Cognitive Biases – Problem Solving

Table des matières

Cette thèse a été préparée à l'Institut des Sciences Cognitives Marc Jeannerod - CNRS
67, Boulevard Pinel - 69675 Bron

Remerciements

Mes remerciements s'adressent avant tout à Hugo Mercier, dont je suis particulièrement honoré d'être le premier doctorant. Sa compréhension de l'humain, sa profonde gentillesse et son extraordinaire efficacité ont fait de moi un des très rares doctorants qui ne s'est jamais plaint de son directeur.

Je tiens d'abord à remercier toutes les personnes qui ont permis à cette thèse d'avoir lieu et donc en premier lieu à la Direction Générale de l'Armement pour l'avoir financée. Je remercie en particulier Didier Bazalgette pour sa confiance et son suivi du projet. Je remercie également Jean-Baptiste Van der Henst pour son accueil chaleureux dans le laboratoire.

Je tiens ensuite à remercier ceux qui ont rendu le long chemin de la thèse plus agréable au quotidien. En particulier un immense merci à Gustavo Estivalet pour la passion qui l'anime dans la vie comme dans la science et à Thomas Castelain pour son incroyable esprit d'équipe et sa grande générosité.

Merci à toutes les personnes qui ont pleinement participé à cette thèse en rendant ces 3 ans plus riches en réflexions comme en éclats de rire. En particulier Timothée Behra, Ludovic Benistant, Laure Bottemane, Audrey Breton, Laurent Cordonnier, Aurianne Couderc, Marie Dekerle, Yang Hu, Romain Mathieu, Sonia Marin, Bruno Martin, Diana Mazzarella, Helena Miton, Quentin Moreau, Ira Noveck, Nicolas Peron, Elisa Pont, Flora Schwartz, Mehdi Senoussi, Pierre Wydoodt.

Merci à tous les membres de l'Institut des Sciences Cognitives Marc Jeannerod qui contribuent à en faire un lieu scientifique rempli d'humanité. En particulier Johan et Sylvain du service informatique pour leur savoir-faire et leur sympathie. Merci également à Christiane Battoue et à Evelyne Robin pour leur précieuse aide dans chaque démarche administrative.

Enfin merci a Georgina Denis pour son aide et son soutien pour l'obtention de cette bourse de thèse. Merci a Nora Parren pour m'avoir supporté dans la période de rédaction et pour son soutien décisif dans la recherche de post-doctorat.

Merci à mes parents pour la profonde confiance qu'ils ont su me donner et pour le soutien inconditionnel qu'ils m'ont apporté tout au long de mes études.

Chapitre 1 - Introduction

De façon aussi indiscutable que l'humain est doué de parole, de mémoire ou qu'il perçoit les couleurs, l'espèce humaine raisonne. Le raisonnement est même souvent considéré comme l'une de ces nobles fonctions cognitives qui nous séparent de l'animal. Pourtant aujourd'hui encore, peu de psychologues s'accordent sur ce qu'est le raisonnement humain, sur ce qui pourrait le caractériser, sur les moyens d'évaluer ses performances, sur ce qui pourrait délimiter son étude, et encore moins sur les mécanismes biologiques qui le sous-tendent. Si, historiquement l'étude du raisonnement est une des disciplines fondatrices des sciences cognitives au même titre que la perception, l'attention ou la mémoire, dans le cas du raisonnement, on peut admettre que les avancées scientifiques des 40 dernières années ont été moins spectaculaires. Certains chercheurs finissant même par douter que le terme de raisonnement soit une catégorie pertinente pour l'étude de la psychologie humaine.

Dans ce chapitre d'introduction nous ne ferons que présenter quelques notions essentielles de notre approche scientifique. Nous présenterons les principes essentiels de la psychologie évolutionniste et entamerons une discussion sur la possibilité que les mécanismes du raisonnement soient issus de la sélection naturelle.

1.1 La pensée adaptative

« Rien en biologie n'a de sens, si ce n 'est à la lumière de l'évolution »

Theodosius Dobzhansky, 1973

La psychologie évolutionniste est née dans les années 90 sous l'impulsion de deux révolutions scientifiques. L'une est la révolution cognitive dans les années 1950-60, fournissant des explications de la pensée en termes d'information et de computation. L'autre est la révolution de la biologie de l'évolution, expliquant la structure complexe du vivant en termes de sélection parmi des réplicateurs. La psychologie évolutionniste n'est pas une discipline mais plutôt une position théorique qui consiste à penser que certaines de nos capacités mentales ont évoluées graduellement au cours de l'évolution pour répondre à des problèmes posés par notre environnement ancestral. Il n'y a en effet, a priori, aucune raison pour considérer le système nerveux central, dont le cerveau fait partie, comme une entité biologique distincte des autres. De la même façon que nos organes sont le fruit de pressions de sélection notre cerveau a lui aussi toutes les chances d'avoir été façonné par la sélection naturelle.

La pensée évolutionniste en psychologie cognitive consiste à former des hypothèses sur la cognition à partir de la correspondance entre la deux niveaux d'explications d'un mécanisme mental : d'une part des explications ultimes, considérant la fonction qu'un mécanisme a pu avoir dans notre environnement ancestral, et d'autre part des explications proximales, ou algorithmiques. A quoi sert cette structure ? Et quel calcul fait cette structure ?

La pensée adaptative en psychologie est la démarche de former des hypothèses sur la structure d'un mécanisme à partir de sa fonction. L'inférence inverse, consistant à faire des hypothèses sur la fonction à partir de la structure, est malheureusement plus souvent observée ; elle correspond à ce que l'on appelle de l'ingénierie à rebours. Si on peut, d'une certaine façon, saluer la recherche de cohérence avec la théorie de l'évolution à partir de théories algorithmiques, on peut aussi noter qu'historiquement la pensée adaptative a, elle, permis de nombreuses avancées dans le domaine de la biologie. Comme le résume le biologiste Ernst Mayr mentionnant l'exemple de William Harvey, médecin du XVIIème siècle :

« The adaptationist question, "What is the function of a given structure or organ?" has been for centuries the basis for every advance in physiology. If it had not been for the adaptationist program, we probably would still not yet know the functions of thymus, spleen, pituitary, and pineal. Harvey's question "Why are there valves in the veins?" was a major stepping stone in his discovery of the circulation of blood. »

Ernst Mayr (1983)

1.2 Modularité de l'esprit

La thèse de la modularité de l'esprit est l'idée que l'esprit est une articulation de différents mécanismes spécialisés. Le fait que les psychologues évolutionnistes soient généralement partisans de cette thèse est loin d'être une coïncidence. Il existe en effet de bons arguments pour défendre l'idée qu'un système modulaire soit optimal pour permettre l'intégration de nouveaux mécanismes neuronaux au fil de l'évolution. Un des arguments repose notamment sur la supériorité des architectures modulaires pour gérer la complexité croissante d'un programme ou d'un logiciel informatique (voir par exemple Carruthers 2005 pour un argumentaire complet en faveur de la modularité massive de l'esprit).

Nous nous inscrivons dans ce cadre de la modularité (massive) de l'esprit, mais précisons toutefois que le terme de module est employé au sens large. Un module est une composante fonctionnelle d'un système mais n'est pas forcément hérité de l'évolution. Après suffisamment d'entrainement nous avons, par exemple, des modules pour l'arithmétique, pour la lecture ou pour les échecs. Le chien de Pavlov a également un module lui faisant activer la représentation de nourriture à chaque son de cloche.

Cette définition de module correspond à la définition biologique ou informatique du terme, mais tranche très clairement avec la notion de module développée par Fodor dans le domaine des sciences cognitives, sa définition contenant bien plus de critères, comme celui de l'innéisme par exemple. Notre définition de module mental se limite à un mécanisme psychologique définit par une fonction algorithmique précise (input, traitement et output bien spécifique). Nous pouvons ensuite ajouter une distinction entre des modules perceptuels, dont l'entrée est sensorielle, et les modules inférentiels dont l'entrée est une représentation. Dans le cas de la perception visuelle, des modules inférentiels intègrent la production de différents modules perceptuels spécialisés dans l'orientation ou la forme des objets par exemple. Le produit de cette intégration pourra ensuite être traité par d'autres modules inférentiels spécialisés dans la reconnaissance d'objets, ou encore la lecture d'émotions sur un visage.

Un module est donc défini par sa description algorithmique (ses entrées, son traitement, ses sorties), mais certains modules ont également une fonction évolutionniste. Le fait de détecter de façon extrêmement rapide et automatique la présence de serpents dans notre environnement par exemple est un exemple classique de module inférentiel utilisant le produit de modules perceptuels dont on peut très facilement imaginer les avantages évolutionnistes qu'il apporte.

Le projet évolutionniste, au sein des sciences cognitives, pourrait se résumer par la recherche des modules hérités de l'évolution, avec l'ambition théorique d'arriver à expliquer avec eux la cognition humaine à partir de recombinaison provenant de l'environnement. Loin, donc, de nier l'importance des mécanismes d'apprentissages, ce sont justement ces mécanismes qui utilisent des unités fonctionnelles non-héritées de l'évolution ; soit simplement associationiste comme le chien de Pavlov, soit comme un mathématicien pensant immédiatement à poser une équation face à un problème numérique. Certaine recombinaisons peuvent être cependant plus profondes, comme dans le cas de l'apprentissage des nombres et de l'arithmétique. Les travaux de Dehaene et collaborateurs, en particulier, proposent un modèle assez complet de ce qu'ils appellent un recyclage neuronal. Par exemple des mécanismes anciens comme les capacités d'estimation numérique, présentes chez les singes, servent de base aux connexions nous permettant de finir par

trouver l'arithmétique intuitive, connexions que les élèves et les professeurs des écoles créeront à la sueur de leurs fronts.

Ces mécanismes d'apprentissage sont d'une importance capitale pour le projet évolutionniste de la cognition car ce sont eux qui permettront de passer des modules quasi-universels, hérités de l'évolution, à toute la complexité humaine dans sa diversité, qu'elle soit culturelle, individuelle, d'un jour à l'autre ou d'une situation à une autre. Historiquement la recherche d'unités fonctionnelles héritées de l'évolution ne fait que commencer et on peut espérer que ce projet de naturalisation de l'esprit humain arrivera un jour à rendre compte de toutes les capacités cognitives spécifiques à l'espèce humaine.

Qu'en est-il du raisonnement ? Le raisonnement est, quasiment par définition, une capacité centrale, très générale, pouvant agir sur une grande diversité de représentation mentale. Si le raisonnement est trop général pour être un module, on ne peut, à priori, envisager que deux conclusions.

Soit le raisonnement n'est pas une capacité modulaire. C'est ce que défendent certains psychologues (par ex. C. Heyes) ou philosophes (par ex. K. Sterelny) qui voient dans la généralité du raisonnement humain une limite à la thèse de la modularité de l'esprit.

Soit, comme le concluent les fondateurs de la psychologie évolutionniste (Tooby et Cosmides 1992), le raisonnement conçu comme une capacité générale n'existerait pas. Ce serait une mauvaise catégorie scientifique qui regrouperait, en fait, différents modules.

Cette thèse défend l'idée que le raisonnement est un mécanisme modulaire malgré son apparente généralité. Nous exposerons, dans le chapitre 4, la solution proposée par Dan Sperber et Hugo Mercier à cet apparent paradoxe. Avant cela nous allons présenter les théories classiques du raisonnement regroupées sous le terme de théories intellectualistes. Nous discuterons, dans le chapitre qui suit, de l'hypothèse logiciste du raisonnement, puis nous passerons aux théories aujourd'hui dominantes dans le domaine de la psychologie du raisonnement : les théories à processus duels.

Chapitre 2 - Logique et raisonnement

2.1 Une histoire de logique

Qu'entend-on exactement par hypothèse logiciste du raisonnement ? Dans sa version précédant la révolution cognitive, c'est l'idée que la logique est la base du raisonnement humain (Henle, 1962). L'influence de Jean Piaget dans la psychologie était alors grandissante. Piaget intégra la tradition logiciste à sa théorie du développement, proposant alors l'idée qu'une fois arrivé à l'âge adulte notre pensée serait guidée par des schémas abstraits et logiques, instanciés à des cas particuliers (Inhelder & Piaget, 1958). Cette idée donnera naissance, avec l'émergence des sciences cognitives, à différentes théories proposant différentes implémentations de lois logiques dans l'esprit humain. Avant de présenter brièvement quelques éléments de ces théories, illustrons la possibilité que de telles structures abstraites puissent guider le raisonnement humain en présentant quelques exemples classiques du paradigme de la déduction.

Pour de bonnes raisons, l'exemple qui ouvre la plupart des ouvrages universitaires sur le raisonnement est un syllogisme catégoriel. En effet, d'une part, un des premiers penseurs à s'essayer à la question du raisonnement – Aristote – utilisait déjà ces problèmes pour différencier les bonnes formes de raisonnement des mauvaises. D'autre part, les syllogismes catégoriels ont sans doute été le type de tâches le plus utilisé dans le domaine, tant ils offrent de nombreuses combinaisons logiques et possibilités d'habillages de leur structure logique, pouvant donner lieu à autant d'expériences. Comme un symbole de l'avancement des idées plutôt lent dans le domaine du raisonnement, ces mêmes tâches se trouvent encore aujourd'hui au centre des débats les plus actuels, comme nous le verrons dans le chapitre suivant.

Prenons un exemple de syllogisme catégoriel d'Aristote :

Prémisses : (1) Tous les humains sont mortels

(2) Tous les Grecs sont humains

Si vous acceptez les énoncés (1) et (2), vous en déduisez que « tous les grecs sont mortels ».

Pour les théories logicistes vous avez utilisé un schéma du type: « Si tous les A sont des B et tous les C sont des A, alors tous les C sont des B ». Les prémisses (1) et (2) étant des cas particuliers de cette règle, vous inférez que « Tous les C (Grecs) sont B (Mortels) ». La logique mentale supposée du raisonnement sert, non seulement à atteindre des conclusions valides en instanciant des règles abstraites sur des prémisses, mais sert également à évaluer la validité de conclusion. Prenons un autre exemple de syllogisme catégoriel mais cette fois dans le but de réaliser l'évaluation et non la production d'une conclusion:

Prémisses :

(2) Aucun chien policier n'est méchant.

(1) Certains chiens dressés sont méchants

Question : Peut-on en déduire que certains chiens dressés ne sont pas des chiens policiers ?

Comme 89% de sujets (Evans 1983), vous répondez oui à la question. Les prémisses permettent bien d'arriver à la conclusion. Nous semblons donc capables à la fois de réaliser des déductions valides et d'évaluer la validité de déduction.

Au delà des syllogismes, nous sommes aussi capables d'utiliser d'autres lois de la logique. Prenons un exemple de disjonction :

Prémisses :

(1) Si je rate mes examens, je vais partir 3 jours en vacances.

(2) Si je réussi mes examens, je vais partir 3 jours en vacances.

Vous déduisez sans effort que cet étudiant va à partir en vacances 3 jours. Une instanciation de la règle de disjonction : « Si A alors B et Si non-A alors B » vous permet de conclure que B se produira à coup sûr.

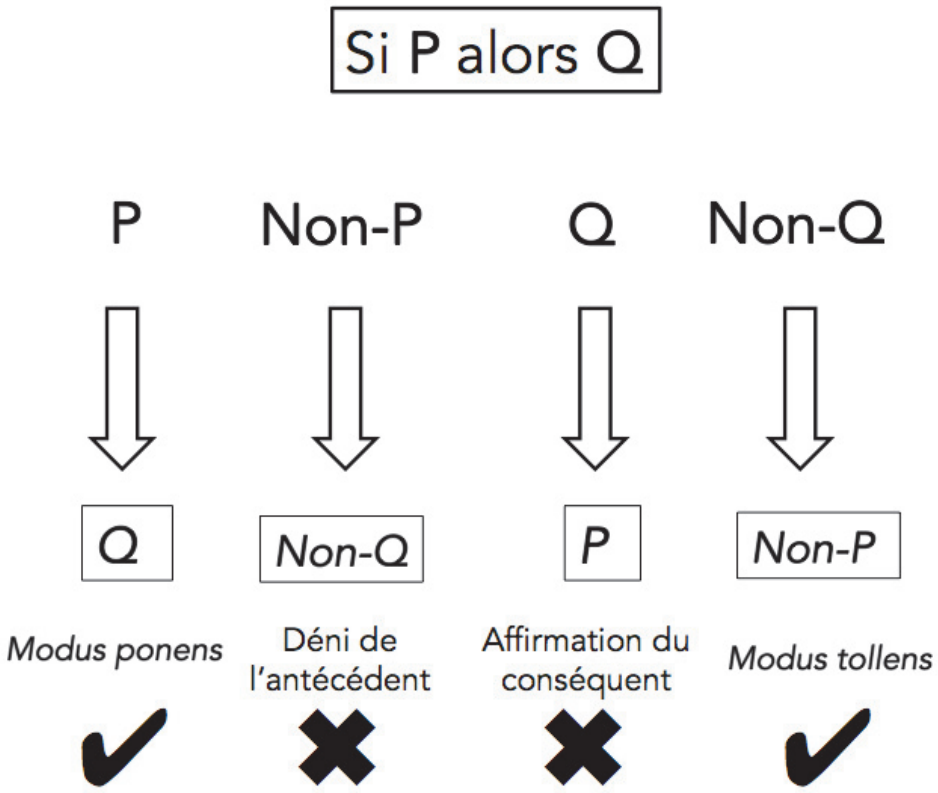
Enfin, donnons l'exemple le plus élémentaire de ce qu'on pourrait attendre d'un raisonnement logique : le raisonnement conditionnel « Si… alors ». Considérons l'exemple suivant :

Prémisses:

(1) Si Marie a un devoir à finir, alors elle restera tard à la bibliothèque.

(2) Marie a un devoir à finir

Vous en déduisez sans problème que Marie restera tard à la bibliothèque. Ce type d'inférence valide se nomme *modus ponens*. Si la prémisse (2) avait été « Marie ne restera pas tard à la bibliothèque » vous auriez pu conclure que « Marie n'a pas de devoir à finir ». Ce type d'inférence, valide également, se nomme *modus tollens*. En revanche si la prémisse (2) avait été « Marie n'a pas de devoir à finir » conclure que « Marie ne restera pas tard à la bibliothèque » est une erreur logique que l'on nomme déni de l'antécédent. Dernier cas, si la prémisse (2) avait était « Marie restera tard à la bibliothèque », ce serait également une erreur de conclure que « Marie a un devoir a finir ». Résumons les inférences valides et non valides d'un raisonnement propositionnel dans la figure suivante :



Les théories logicistes voient donc le raisonnement comme produisant et évaluant des arguments logiques, cette capacité nous permettrait d'atteindre des conclusions valides (ou de vérifier leur validité) lorsque certaines prémisses sont traitées. Comment notre esprit réalise-t-il cela d'un point de vue algorithmique? Deux théories se sont affrontées pendant toute la dernière partie du XXième siècle sur cette question.

D'un côté plusieurs chercheurs, dans la lignée de Piaget - en particulier Martin Braine (1978, Braine & O'Brien,1998) et Lance Rips (e.g. 1983, 1994)-, ont défendu l'idée que nous arrivons à réaliser des déductions logiques grâce à une « logique mentale ». Notre raisonnement fonctionnerait de façon comparable à des preuves logiques, utilisant des règles inférentielles abstraites intégrées à notre esprit.

A partir de 1983, Philip Johnson-Laird va proposer une théorie qui se veut radicalement différente de la logique mentale. Plutôt que d'appliquer des règles abstraites déjà présentes à l'esprit, cette théorie alternative propose que nous représentions et intégrions le contenu des prémisses en construisant des « modèles mentaux » à partir des prémisses, comparables à des images mentales schématiques de la situation. Ce que ferait

notre raisonnement serait d'envisager des mondes possibles et, si aucun monde dans lequel la conclusion n'est pas vraie n'est trouvé, l'argument est valide. Si, à l'origine, cette théorie était présentée comme une critique du logicisme, à bien des égards la théorie des schémas mentaux ne s'est pas vraiment démarquée de ce courant, ne serait-ce que par le fait que cette théorie reste focalisée sur le paradigme de la déduction logique.

Ce qui distingue le plus clairement la théorie de la logique mentale de la théorie des modèles mentaux, c'est ce qui rend une tâche de raisonnement difficile. Pour les partisans de la logique mentale, c'est le nombre de règles et d'étapes qui doivent être suivies. Pour les partisans des modèles mentaux, c'est le nombre de modèles qui doivent être construits et intégrés pour arriver à une conclusion donnée. Pour le reste, ces deux théories partagent l'idée que l'humain a des mécanismes capables de produire des inférences logiques valides. Elles partagent aussi l'idée que l'humain utilise, d'une façon ou d'une autre avec plus ou moins de succès, la logique pour assurer sa rationalité. Citons un passage de (Oaksford & Chater 1995) à propos de ces deux théories :

« La logique mentale et les modèles mentaux prennent la logique comme une théorie au niveau computationnel. C'est évident pour l'approche de la logique mentale. Mais cela découle aussi immédiatement dans le cas de l'approche des schémas mentaux, étant donné que le but de la théorie des schémas mentaux est de fournir un mécanisme qui réaliserait, de façon logique, des inférences déductives valides » (p. 133)

Quoiqu'il en soit, au delà de cette bataille interne, le principal problème que va rencontrer l'hypothèse logiciste, dès son commencement, c'est que les données expérimentales ne vont pas dans son sens. Les performances humaines en logique sont désormais reconnues, de façon consensuelle, comme globalement mauvaises. Nous allons le voir, de nombreuses études expérimentales ont démontré que nos performances en raisonnement logique son, non seulement mauvaises dans l'abstrait, mais dépendent également du contexte dans lequel la structure logique est incarnée. Ce qui représente des observations pour le moins problématiques pour des mécanismes censés être abstraits, généraux et basés sur la logique.

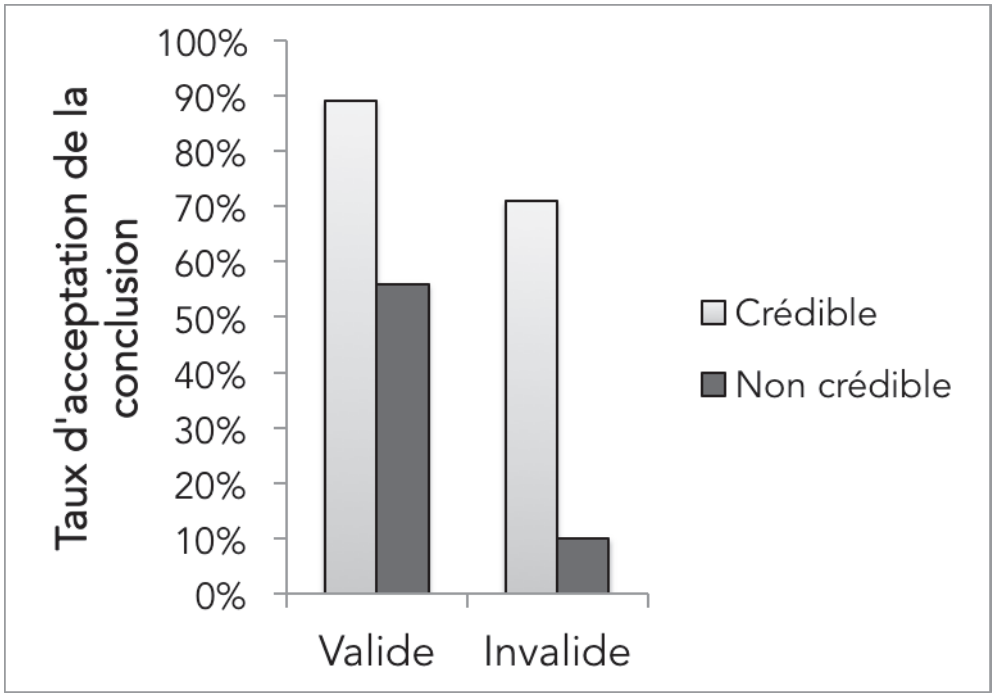
2.2 Une logique remise en question

Entre les années 60 et le début des années 2000, quarante années de paradigme de la déduction ont donné lieu à une expérimentation massive des compétences logiques humaines. Si les conclusions à tirer restent en discussion, le constat est sans appel : les performances humaines en logique sont, pour le moins, médiocres.

Comme le résume Evans dans un article faisant le bilan du paradigme de la déduction, dont il a lui-même été un acteur majeur: « ... les performances logiques des sujet dans des tâches de raisonnement abstrait sont généralement assez mauvaises » (Evans 2002). D'une part les sujets acceptent fréquemment des conclusions non-valides, que ce soit dans des syllogismes ou dans le raisonnement conditionnel abstrait (Evans, Newtsead & Byrne 1993). Mais toutes les bonnes inférences sont également loin d'être réalisées, en particulier dans le cas du raisonnement conditionnel abstrait, si le *modus ponens* est massivement réalisé par les participants, le *modus tollens* lui n'est produit que dans environ 60% des sujets et pas dans n'importe quelle population : chez des étudiants d'université. Ce constat pourrait être perçu comme un argument majeur à l'encontre d'une vision logiciste. N'est-ce pas paradoxal qu'un mécanisme traitant de logique soit si peu performant dans l'abstrait ? Peut-être pas. Après tout, que le raisonnement, dans sa vision logiciste, soit un mécanisme né de pressions de sélection (aussi difficiles à imaginer puissent-elles être) ou un mécanisme d'apprentissage captant les structures logiques de l'environnement, les mécanismes du raisonnement auraient été, de toute façon, baignés dans des situations concrètes. Notre pensée pourrait parfaitement utiliser des lois internes semblables à celles de la logique sans pour autant être performante avec les formes abstraites elles-mêmes. On voit assez mal quels problèmes aux contenus abstraits pourraient être résolus par une pensée logique, tant ces stimuli sont peu présents, au cours de l'évolution de notre espèce bien sûr, mais également dans la vie d'un humain moderne dont la logique n'est pas le métier. Le raisonnement n'est pas à l'aise dans l'abstrait, soit. Son rôle serait plutôt d'instancier des règles abstraites, quel que soit le contexte, à condition qu'il y ait un contexte. Malheureusement pour les théories logicistes, le raisonnement est loin d'être indépendant du contexte.

Reprenons les syllogismes à propos des chiens policiers dont vous aviez, comme la plupart des sujets, réussi à établir la validité. Prenons à présent des syllogismes de la même forme logique mais en faisant varier, soit la crédibilité des prémisses, soit la validité de la conclusion comme l'illustrent les exemples ci-dessous, tirés d'Evans 1983

	VALIDE	INVALIDE
CREDIBLE	Aucun chien policier n'est méchant Certain chiens dressés sont méchants Donc, **certains chiens dressés ne sont pas des chiens policiers**	Aucun produit addictif n'est bon marché Certaines cigarettes sont bon marché. Donc, **certains produits addictifs ne sont pas des cigarettes**
NON CREDIBLE	Aucune chose nutritive n'est bon marché Certains comprimés vitaminés ne sont pas bon marché Donc, **certains comprimés vitaminés ne sont pas nutritifs**	Aucun milliardaire n'est un dur travailleur Certaines personnes fortunées sont des durs travailleurs Donc, **certains milliardaires ne sont pas des personnes fortunées**



Les performances des sujets, lorsqu'ils résolvent les syllogismes présentés plus haut, nous indiquent que le raisonnement logique est sensible aux croyances initiales et au contexte. Pour la vision logiciste ce résultat est interprété comme un biais, au sens où nos croyances sur le monde viennent perturber le bon déroulement d'une pensée logique. Notre raisonnement serait parvenu à déduire et reconnaître les bons arguments logiques si des éléments contextuels ou le manque de ressources cognitives (typiquement la mémoire de travail) ne l'avaient pas gêné dans son fonctionnement.

Les plus éminents psychologues de la psychologie du raisonnement ont focalisé leur attention sur les performances humaines à résoudre les syllogismes catégoriels d'Aristote. Bien sûr, en science, l'étude de phénomènes marginaux ou sans importance pratique peut être d'une grande pertinence théorique, mais cela est difficilement défendable dans le cas de l'étude des syllogismes. Dans un article de synthèse publié en 2012, après un demi-siècle d'études intensives, Sange, Khemlani et Philip Johnson-Laird identifient douze théories concurrentes du raisonnement syllogistique, dont aucune, disent-ils, « ne rend compte des résultats de façon adéquate». «L'existence de 12 théories de tout domaine scientifique », ajoutent-ils," est un petit désastre." Cette accusation est d'autant plus

remarquable que l'une des douze théories, et sans doute l'une des plus influentes, est la théorie des modèles mentaux de Johnson-Laird lui-même.

Quoiqu'il en soit, l'addition de contexte à des tâches abstraites ne facilite donc pas forcément les performances logiques humaines. Mais il est vrai que, même dans ces cas, le contexte reste très artificiel. Les sujets venant passer une expérience sur des syllogismes ne se demandaient probablement pas si tous les chiens dressés étaient des chiens policiers et encore moins si certains milliardaires n'étaient pas fortunés. L'artificialité des tâches, l'ennui qu'elles peuvent procurer, sont sans doute des facteurs qui ne participent pas au déploiement des mécanismes de raisonnement humain, quels qu'ils soient.

De façon parallèle au paradigme de déduction logique dont les résultats de non-normativité agitaient les psychologues du raisonnement en Europe, Outre-Atlantique d'autres chercheurs ont réalisé des observations similaires dans le domaine plus général du jugement et de la prise de décision (principalement économique). Daniel Kahneman et Amos Tversky ont réalisé, dans les années 70 et 80, une série d'expériences démontrant le non-respect de règles normatives lorsque les humains prennent des décisions. Contrairement au paradigme de la déduction, la question de savoir si, ou comment, notre esprit utilise la logique propositionnelle est très éloigné des intérêts de ces chercheurs. Les lois normatives sur lesquelles ils vont concentrer leur travaux n'est pas la logique mais la théorie des probabilités et la théorie rationnelle du choix. Si on ne peut pas vraiment situer les travaux de Kahneman et Tversky dans le courant logiciste du raisonnement, il faut remarquer qu'ils font aussi l'hypothèse que le respect de lois normatives est le modèle standard : tout écart à une loi normative se doit d'être expliqué par les psychologues. Une entreprise que ces deux chercheurs ont passé leur carrière à réaliser, en terme d'heuristiques et de biais, avec un succès mondial. Parmi les nombreuses tâches mettant en évidence la violation de loi normative dans la prise de décision, celle utilisée dans (Shafir & Tversky, 1992; Tversky & Shafir, 1992b) illustre une violation de la loi de disjonction que nous avons vue précédemment. L'étudiant qui part en vacances ayant réussi ou non ses examens.

L'expérience de Shafir et Tversky est la suivante.

Tout d'abord deux groupes de sujets reçoivent le texte ci-dessous, soit dans sa version « réussite » des examens (en italique), soit dans sa version « échec » (en gras).

Imaginez que vous veniez juste de passer une série d'examen particulièrement éprouvant. C'est la fin du semestre, vous vous sentez très fatigué et vous venez d'apprendre que vous avez
réussi les examens.
[OU]
échoué les examens, vous devrez repasser ces examens dans quelques mois –après les vacances de Noël.

Vous avez l'opportunité d'acheter un attrayant voyage de 5 jours tout compris à Hawaï à un prix remarquablement bas. L'offre en question expire demain. Quelle option choisissez-vous ?

(a) Souscrire à l'offre de vacances. *54%* / **57%**

(b) Ne pas souscrire à l'offre de vacances.*16%* / **12%**

(c) Payer un supplément de 5€ non-remboursable pour pouvoir souscrire à l'offre après-demain. *30%* / **31%**

Comme les pourcentages de réponses données par les sujets l'indiquent, que les sujets pensent avoir réussi ou raté leur examens ne change quasiment pas la distribution des réponses. La majorité de sujets choisissent de souscrire à l'offre.

Observons à présent les résultats d'un troisième groupe de sujets. Dans ce cas le scenario laisse ouverte la possibilité d'avoir réussi ou échoué aux examens :

Imaginez que vous veniez juste de passer une série d'examens particulièrement éprouvants. C'est la fin du semestre, vous vous sentez très fatigué et vous n'êtes pas sûr d'avoir réussi les examens. Dans le cas où vous auriez échoué, vous auriez à repasser ces examens dans quelques mois –après les vacances de Noël. Vous avez l'opportunité d'acheter un attrayant voyage de 5 jours tout compris à Hawaï à un prix remarquablement bas. L'offre en question expire demain, alors que les notes des examens ne seront disponibles qu'après-demain. Quelle option choisissez-vous ?

(a) Souscrire à l'offre de vacances. **32%**

(b) Ne pas souscrire à l'offre de vacances. **7%**

(c) Payer un supplément de 5€ non-remboursable pour pouvoir souscrire à l'offre après-demain, après avoir eu les résultats des examens. **61%**

On remarque alors que plus de participants décident de partir en vacances lorsqu'ils connaissent le résultat des examens que lorsqu'ils ne le connaissent pas, et ce quel que soit le résultat des examens.

Les auteurs de cette étude présentent ces résultats comme une violation des normes de la théorie de la décision plutôt que comme violation de la logique propositionnelle, mais c'est la même règle exprimée dans deux théories normatives différentes. Cet exemple illustre le fait que même des règles en apparence aussi triviales que la disjonction ne sont pas forcément réalisées aussi facilement dans tous les contextes.

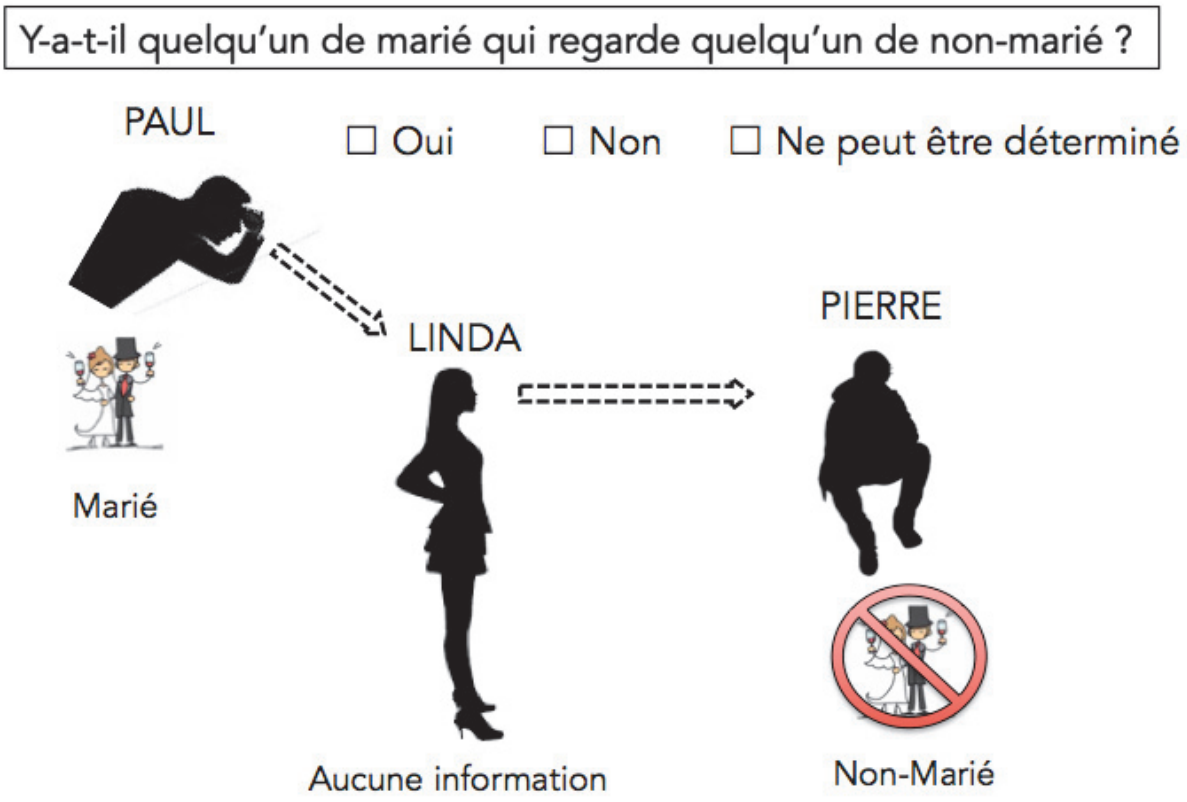
Un exemple encore plus frappant de contexte où la disjonction n'est pas réalisée est proposé par Levesque en 1986. La tâche, illustrée dans la figure ci-dessous, est présentée aux sujets sous la forme suivante :

Paul regarde Linda. Linda regarde Pierre

Paul est marié. Pierre n'est pas marié.

Y-a-t-il quelqu'un de marié qui regarde quelqu'un de non-marié ?

□ Oui
□ Non
□ Ne peut être déterminé



Comme le montre Stanovich & Toplak (2002) la grande majorité des sujets répond « on ne peut pas savoir » à la question, alors que la bonne réponse est « oui » (13% de réussite). Les sujets ne réalisent pas la disjonction nécessaire pour arriver à la bonne réponse : « Si Linda est mariée, elle regarde Pierre, non-marié, donc la règle fonctionne. Si Linda est non-mariée, Paul, qui est marié, la regarde. Donc la règle fonctionne dans les deux cas ».

La structure de disjonction qui semble si évidente dans le premier exemple présenté est donc largement non réalisée dans d'autres contextes. Bien sûr, dans ce cas également, on peut défendre que cela soit dû à un manque de ressources ou que le contexte perturbe la bonne exécution de la règle de disjonction. Cette tâche fait partie de notre matériel expérimental, ce sera pour nous la tâche de « Paul et Linda ».

De façon encore plus problématique pour les théories logicistes, même pour la structure logique qui reste, de loin, la moins ratée par les sujets -le *modus ponens*-, il ne semble pas non plus très difficile de perturber la production de cette inférence logique.

Les sujets n'ont aucun mal à produire un *modus ponens* dans l'exemple précédent de Marie qui restera tard à la bibliothèque si elle a un devoir à finir. Ruth Byrne en 1989 donne à des sujets le même énoncé mais en ajoutant une prémisse par rapport à l'exemple précédent (soulignée ici).

Prémisses:

(1) Si Marie a un devoir à finir, alors elle restera tard à la bibliothèque.

(1') Si la librairie reste ouverte, alors Marie restera tard à la bibliothèque

(2) Marie a un devoir à finir

D'un point de vue strictement logique, comme vous l'avez fait dans l'exemple similaire sans la prémisse (1'), vous devriez déduire que « Marie restera tard à la bibliothèque. ». Pourtant dans cette version seulement 38% des sujets produisent ici un *modus ponens*.

Si, comme les partisans de la logique mentale le prétendent, les gens avaient une règle mentale du type *modus ponens*, alors l'inférence devrait être automatique, quel que soit le contexte. Les participants sont chargés de prendre les prémisses comme vraies, donc, compte tenu des prémisses, «Si Marie a un devoir à finir, elle va étudier tard dans la bibliothèque » et «Marie a un devoir à finir », ils devraient infailliblement conclure qu'elle restera tard à la bibliothèque. Qu'en est-il de la possibilité que la bibliothèque soit fermée? Marie pourrait très bien avoir un laisser-passer pour travailler dans la bibliothèque quand elle est fermée. Mais ces considérations sont sans importance pour une tâche logique, tout comme le fait de penser qu'un chocolat pourrait fondre est sans rapport avec la tâche arithmétique d'ajouter trois chocolats à deux chocolats.

Les partisans de la logique mentale ont cependant suggéré une proposition alternative pour expliquer ce résultat potentiellement embarrassant. Il est possible, par exemple, de regrouper les deux prémisses (1) et (1') en une seule : «Si Marie a un devoir à finir *et* si la bibliothèque reste ouverte alors Marie va étudier tard dans la bibliothèque." C'est après tout un moyen réaliste de comprendre la situation. Si c'est bien la façon dont

les gens interprètent les prémisses (1) et (1'), la prémisse (2) «Marie a un devoir à finir," ne suffit pas à déclencher l'inférence valide de *modus ponens*. Ce résultat n'est donc pas une preuve contre la logique mentale.

Notons tout de même, à ce stade, que même si les performances humaines sont globalement mauvaises dans les tâches de logique, les partisans des théories logicistes du raisonnement come Evans insistent sur le fait que le paradigme de la déduction a, malgré tout, apporté des résultats indiquant la présence d'une « compétence déductive minimale et irréductible pour laquelle les psychologues doivent fournir une explication ». Il précise : « des participants non-entrainés à la logique formelle sont capables d'établir la validité d'un argument bien au-dessus du hasard » (Evans 2002). Soit, admettons que toute théorie du raisonnement concurrente de la vision logiciste devra expliquer ces performances logiques pour le moins minimales.

Avant de conclure ce chapitre sur l'hypothèse logiciste du raisonnement, présentons la tâche qui fut sans doute la plus discutée dans l'histoire du paradigme logiciste (voir Evans, Newstead & Byrne, 1993 ; Manktelow, 1999 pour une revue de littérature) et qui, avec les syllogismes catégoriels, sera au centre d'un changement tant attendu dans le domaine du raisonnement : l'introduction des théories à double processus.

En 1966, Peter Wason présenta pour la première fois la tâche suivante :

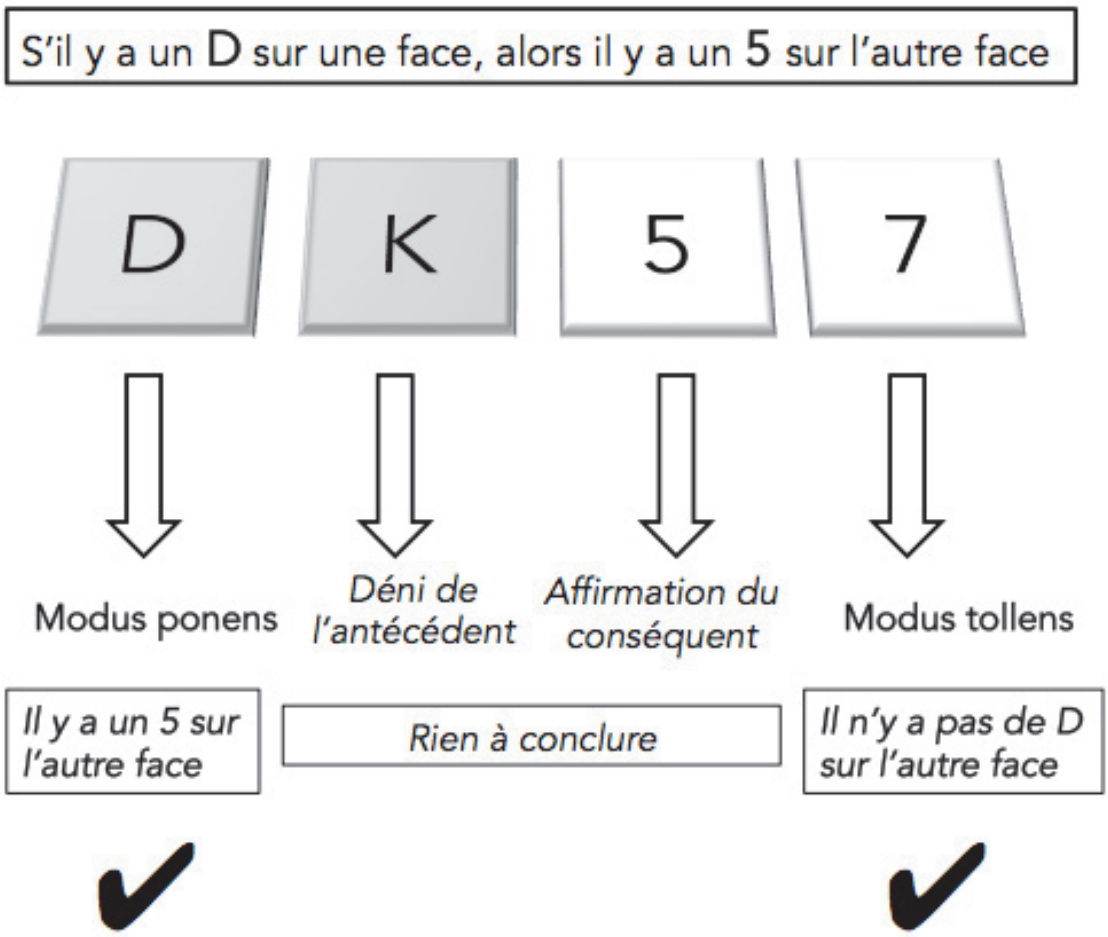
« En face de vous se trouvent quatre cartes »

« Chaque carte a une lettre sur une face et un nombre sur l'autre. Deux cartes (E et K) on la face lettre visible. Deux cartes (2 et 7) ont la face chiffre visible.»

Votre tâche est de répondre à la question suivante :

« Laquelle ou lesquelles de ces quatre cartes est-il nécessaire de retourner pour décider si la règle suivante est vraie ou fausse : 'S'il y a un D sur une face, alors il y a un 5 sur l'autre face' ».

La structure logique de la tâche est la suivante :



Il faut donc sélectionner les cartes D et 7, ce que seul 10% des sujets environ fait. Cet échec massif des sujets à une tâche de logique fut l'un des premiers résultats expérimentaux qui mit en doute l'idée que l'esprit humain utilise des lois logiques. Peter Wason utilisa d'ailleurs cette tâche contre les premières théories logicistes inspirées par Jean Piaget. En 1968 Wason écrit :

« Si Piaget a raison, alors les sujets de cette expérience devraient avoir atteint le stade opératoire formel »… « Est-il possible que le stade opératoire formel ne soit pas complètement atteint à l'adolescence, même pour des individus intelligents ? »

A l'époque ces premiers résultats font conclure des psychologues comme Wason, non pas que la logique n'est pas le bon outil, mais que les humains sont illogiques, et donc irrationnels.

En 1970, Evans proposa une version légèrement différente de cette même tâche. Seul la règle de départ est modifiée : « S'il y a un D sur une face alors il n'y a **pas** de 5 ». Dans ce cas là, la majorité des participants donne la réponse logiquement valide : 5 et D. En fait, dans le deux cas, Evans remarque que la réponse donnée par la majorité des participants correspond aux cartes données dans l'énoncé. Dans le premier cas cela mène à une erreur logique. Dans le deuxième cela mène à une bonne réponse. Ces résultats feront dire à Evans que les sujets ne raisonnent pas vraiment dans la tâche de Wason, ce qui peut paraître surprenant du point de vue logiciste. Mais peut-être mieux vaut-il penser que les sujets ne raisonnent pas plutôt qu'ils échouent à raisonner logiquement. Encore une fois d'autres mécanismes sont désignés comme coupables de la non-activation du raisonnement. Dans ce cas, Evans défend l'idée d'un «biais d'appariement » qui ferait choisir aux sujets les deux cartes énoncées dans l'énoncé. En fait il a plus tard été démontré que les sujets suivent plutôt des principes de pertinence.

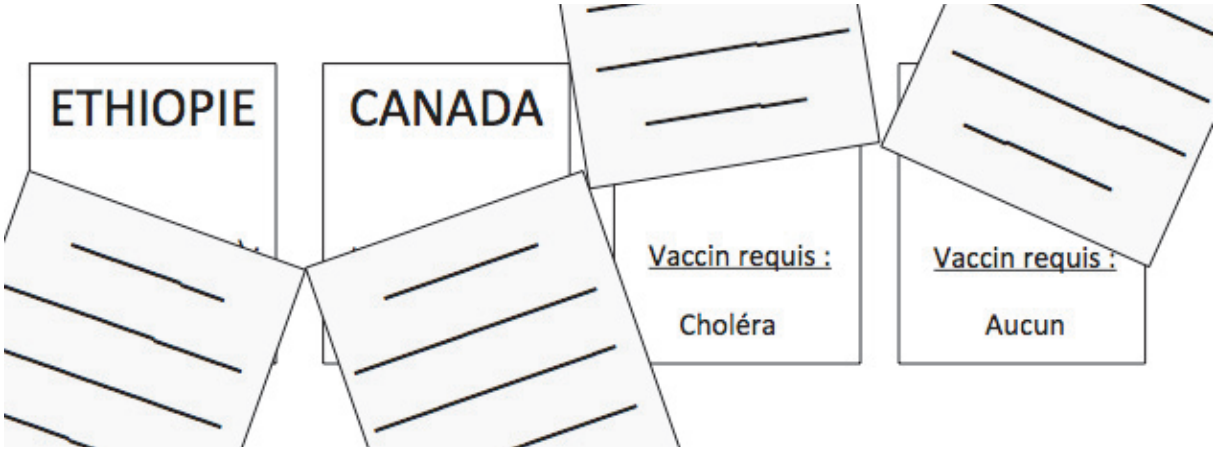
Présentons la dernière étude expérimentale de ce chapitre, que l'on peut considérer comme le travail le plus abouti sur la tâche de Wason.

Girotto Vittorio, Markus Kemmelmeir, Jean-Baptiste Van der Henst & Dan Sperber dans un article de 2001 intitulé "Raisonneurs incompétents ou virtuoses pragmatiques ?" ont fait résoudre aux même participants, la tâche de Wason quatre fois de suite.

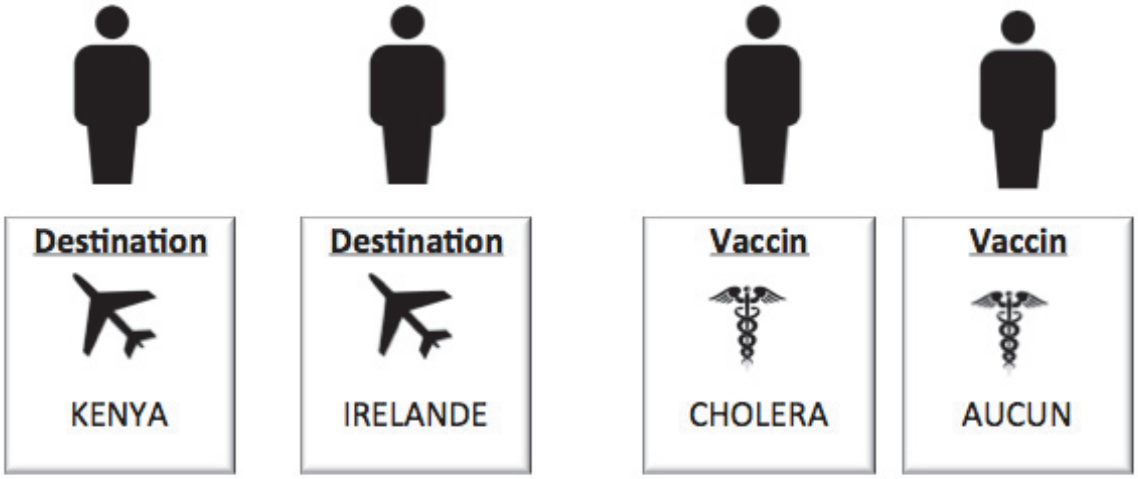
Voici la première version:

Imaginez- vous dans les années 80, vous travaillez pour une agence de voyage et un client vous dit : « je voudrais aller en Afrique de l'Est mais je suis allergique au vaccin contre le choléra ». Vous lui répondez que tout voyage dans un pays d'Afrique de l'Est nécessite un vaccin contre le choléra. Il ne vous croit pas. Afin de le convaincre vous décidez de lui

montrer des documents de l'agence. Votre agence utilise des fiches pour chaque pays : en haut en majuscule se trouve le nom du pays, en bas de la fiche se trouve les vaccins nécessaires pour ce pays. Quatre fiches se trouvent face à vous mais certaines parties sont recouvertes par d'autres documents, voici un schéma de ce que vous voyez :



Dans le deuxième scénario, l'existence de la règle ayant été confirmée, les participants devaient vérifier si quatre clients de l'agence avaient bien obéi à cette règle sur le choléra. Cette fois, les cartes montrent d'un côté un pays auquel le client avait l'intention de voyager, de l'autre côté, les vaccinations de ce client. Voici un schéma :

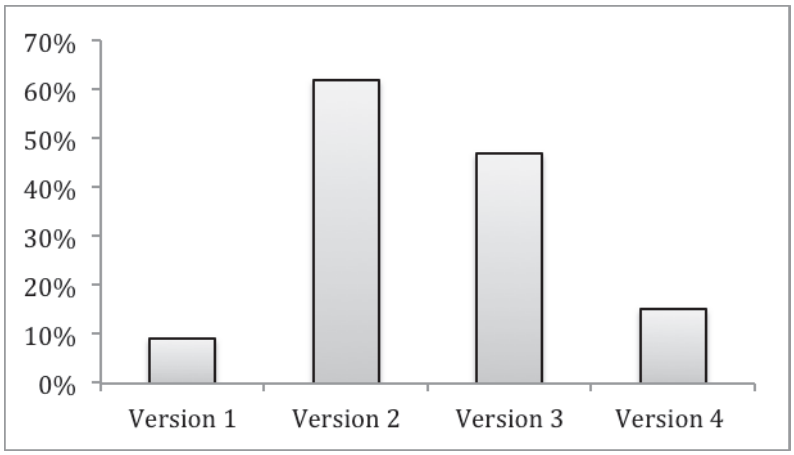


Les troisième et quatrième scénarios sont semblables au premier et au second, mais cette fois l'histoire ne se déroule plus dans les années 80, mais dans le présent.

Dans le troisième scénario, ce n'est plus un client mais vous-même, employé de l'agence, qui souhaitez partir en Afrique de l'Est. Vous êtes allergique au vaccin contre le choléra et pensez que la règle sur le choléra n'est plus en vigueur. Votre patron est en désaccord. La question est de savoir s'il est toujours vrai que la même règle sur le choléra est en vigueur. Cela représente cette fois un doute et non une confirmation à donner à un client comme dans le premier épisode.

Dans le dernier scénario, vous aviez raison de penser que la règle n'est plus en vigueur. Votre patron vous demande maintenant de vérifier si des clients de l'agence ont été induits en erreur en obéissant à cette fausse règle.

On pourrait donc s'attendre à des taux de réussites comparables, mais il n'en est rien. Les performances des sujets sont les suivantes :



Ces performances s'expliquent parfaitement par des considérations pragmatiques en terme de maximisation de pertinence. Ces résultats montrent non pas que les gens sont de mauvais raisonneurs, mais plutôt qu'ils sont des virtuoses de la pragmatique, capables d'ajuster leur interprétation de la tâche à la demande des différentes situations.

2.3 Conclusion sur la pensée logique

Face à l'ensemble des résultats quelque peu accablants pour une vision logiciste du raisonnement, et dont nous n'avons présenté qu'un aperçu, les plus ardents défenseurs d'une logique mentale peuvent avoir deux réactions.

Ils peuvent essayer d'assouplir les lois de la logique. Après tout, nous n'avons jusque là considéré que la logique propositionnelle classique alors que, depuis, la science de la logique a réalisé d'énormes avancées. En particulier, parmi les chercheurs qui proposent d'autres normes pour évaluer les performances du raisonnement humain, on retrouve une logique non-monotone (Stenning et Lambalgen 2008), ou encore une logique probabiliste (Baratgin, Over et Politzer 2014), mais également des modèles bayésiens (Oaksford et Chater 2007).

Ils peuvent également, malgré tout, mettre en avant les réussites des sujets, aussi peu nombreuses soient-elles, et ajouter l'idée que d'autres mécanismes viennent contraindre ou perturber le raisonnement. C'est plutôt cette dernière stratégie que les psychologues du raisonnement ont majoritairement adoptée et que nous allons explorer dans la suite.

Que peut-on tirer de toutes ces années de recherche profondément imprégnées d'une vision logiciste du raisonnement ? Au moins l'idée que l'humain, si tant est qu'il soit pourvu de mécanismes d'inférence logique, a quelques difficultés à les utiliser.

Une autre conclusion est aussi apparue au cours des débats autour de l'hypothèse logiciste : l'humain est irrationnel. Ce serait la grande conclusion à la fois du paradigme de la déduction logique et des travaux sur la prise de décision menés par Kahneman et Tversky. De notre point vue, cependant, dire que l'humain est irrationnel est soit évident, soit faux.

D'un coté cela est évident car l'hypothèse inverse est extrêmement peu plausible. Observer que l'humain viole systématiquement certaines théories normatives comme la logique classique ou la théorie des probabilités n'est surprenant que si l'on suppose ces structures abstraites comme essentielles à la pensée. Or, si cette hypothèse a du sens d'un point de vue de l'histoire des idées, l'observation de déviation par rapport à des lois normatives sophistiquées et dénuées de tout contexte n'est pas vraiment surprenante. Jusqu'à très récemment dans l'histoire de l'humanité, aucun humain n'a eu à respecter ou même à reconnaître, au cours de sa vie, ce type de lois normatives en tant que telles. Que nous fassions des inférences qui enrichissent notre pensée et que nous ayons à maintenir une certaine cohérence est indéniable, mais la logique ou les mathématiques sont loin

d'être les seuls moyens de réaliser cela. S'il s'était avéré que le raisonnement respectait les lois de la logique classique, la théorie du choix et des probabilités, des centaines de milliers d'années avant leur invention, là aurait été la vraie surprise. Notons que le fait que certains mécanismes psychologiques puissent être approximés quasiment parfaitement par des modèles bayésiens (voir par exemple Tenenbaum, Griffiths & Kemp 2006 dans le cas de l'apprentissage catégoriel) représente une victoire scientifique pour une norme à partir de laquelle on peut accepter l'idée de rationalité. Le niveau de prédiction est tel que cette norme apparaît comme synonyme de la stratégie optimale pour ce genre de problème computationnel. On peut difficilement dire la même chose des prédictions des modèles logiques pour les performances du raisonnement. Le fait de considérer l'humain comme irrationnel plutôt que de changer de norme vient sans doute d'une trop grande importance données aux lois logiques dans nos explications psychologiques ce qui, disons-le à nouveau, se comprend mieux d'un point de vue historique. En revanche, cela ne diminue en rien l'intérêt d'étudier les écarts systématiques à ces normes, ou de déterminer les contextes plus ou moins propices à leur violation. Que ce soit dans des domaines où la rationalité des agents est assumée pour des raisons de maniabilité des modèles (comme l'économie ou la modélisation multi-agent par exemple), ou pour des considérations appliquées dans les domaines où ces lois sont pertinentes (par l'exemple pour identifier des freins à l'apprentissage des mathématiques ou encore pour limiter les erreurs des traders ou des physiciens).

D'un autre point de vue, dire que l'humain est irrationnel a toute les chances d'être faux. Dans une perspective évolutionniste de la cognition en particulier, nos mécanismes mentaux ont évolué pour s'adapter à notre environnement. Ces processus sont rationnels au sens où ils réalisent leur fonction. Dans les rares cas où ils échouent à réaliser leur fonction, on peut parler d'irrationalité mais on peut sans doute se permettre de considérer comme des anomalies (ce qu'un bon modèle de ces mécanismes devrait avoir dans leur erreurs de prédiction). Le fait que ces mêmes mécanismes nous fassent violer certaines normes ne les rend pas irrationnels mais simplement non adaptés à la norme en question.

Cette question de la rationalité a donné lieu à de vifs débats entre Gigerrenzer et Kahneman (Kahneman & Tversky 1996, Gigerrenzer 1996) qui ont sans doute participé à un

certain avancement des idées du domaine. Les partisans des théories logicistes vont alors parler de « paradoxe de la rationalité » : d'un côté l'espèce humaine est intelligente, de l'autre les laboratoires de psychologie du raisonnement accumulent les observations d'erreurs de raisonnement. C'est face à ce paradoxe qu'Evans et Over 1996 proposeront une théorie alternative de la rationalité. Il y aurait en fait deux rationalités : une rationalité personnelle (rationalité 1) et une rationalité normative (rationalité 2) (Evans et Over 1996). Ces mêmes auteurs vont aussi parler de deux systèmes cognitifs associés à chacune de ces rationalités. L'ère des théories à processus duels peut commencer.

En conclusion, depuis le commencement du questionnement sur le raisonnement, les penseurs ont cherché à distinguer les bonnes des mauvaises inférences. Il semble qu'une confusion ait ensuite perduré entre, d'un côté l'étude du raisonnement tel qu'il est - comme mécanisme psychologique – et de l'autre l'étude du raisonnement tel qu'il devrait être - comme principes normatifs-. Encore aujourd'hui un bon raisonnement tend à être associé à un raisonnement logique et un mauvais raisonnement à un raisonnement qui a échoué à l'être. Si l'on admet sans mal qu'étudier l'écart entre les décisions humaines et des normes peut être pertinent, dans d'autres disciplines que la psychologie, en revanche, pour ce qui est de l'étude de mécanismes psychologiques, les liens avec la logique sont beaucoup moins clairs. Même si nous n'avons pas apporté ici d'éléments qui forceraient à un divorce définitif entre logique et raisonnement, il apparaît qu'il n'y a pas non plus, a priori, de bonnes raisons pour considérer les inférences logiques comme centrales dans l'étude de la psychologie humaine, ni même simplement comme pertinentes.

L'étude expérimentale du raisonnement et l'observation des piètres performances humaines en logique ont en tout cas déplacé l'attention des chercheurs sur l'interaction entre le raisonnement et d'autres mécanismes. Comme nous allons le voir dans le chapitre suivant, le défi pour les théories qui émergent ensuite est, d'une part, de caractériser les processus du raisonnement, d'autre part d'expliquer comment l'interaction entre le raisonnement et d'autres mécanismes peut rendre compte des performances du raisonnement humain.

Chapitre 3 - Les théories à double processus

3.1 Naissance et domination des théories à double processus

La première esquisse d'une théorie à double processus fut énoncée par Jonathan Evans et Peter Wason, dans deux articles publiés en 1975 et 1976. Ces premiers travaux furent cependant vite oubliés. Comme nous l'avons vu, en ajoutant un «non» dans la règle de la tâche de Wason, Evans a démontré que les sujets font leur choix sans réellement raisonner. Ils choisissent simplement les cartes qu'ils voient intuitivement comme pertinentes. La sélection des cartes est donc basée sur un processus intuitif.

Evans et Wason ont refait l'expérience, en demandant cette fois aux sujets d'expliquer leur choix, s'assurant ainsi que les sujets raisonnent. En effet, si les sujets ne raisonnaient pas pour résoudre le problème - résolu intuitivement – ils raisonnaient au moins pour justifier leur solution intuitive. Lorsque leur solution s'avèrait être logiquement correcte (généralement quand la règle était modifiée), les sujets fournissaient une justification qui semblait logique. Lorsque leur solution était incorrecte en revanche, les sujets donnaient, avec tout autant de confiance, une justification qui n'avait pas de sens logique.

Ce que les processus du raisonnement conscient semblent faire c'est tout simplement fournir une rationalisation pour un choix qui a été fait avant l'implication du raisonnement.

Trois idées nourrissent cette première esquisse d'une théorie à processus duels. La première de constater un contraste classique entre deux modes d'inférence, l'un se produisant spontanément et sans effort, l'autre, le raisonnement à proprement parler – est au contraire délibéré et coûteux en ressources cognitives. Une seconde idée, plus originale, est que les gens approchent la même tâche déductive dans les deux modes. Dans la tâche de Wason par exemple, la plupart des participants produisent à la fois une sélection spontanée de cartes et une explication motivée de leur sélection.

La troisième idée est sans doute la plus provocatrice: ce que les processus de type délibératifs font, c'est simplement rationaliser une conclusion qui avait été obtenue par des processus intuitifs. Cette idée rabaissant le rôle du raisonnement a sans doute participé à l'accueil pour le moins mitigé qu'ont reçu les théories d'Evans et Wason à l'époque.

L'approche processus duels du raisonnement n'a presque jamais été mentionnée et encore moins discutée dans les vingt années qui suivirent.

Quand elle fit sa réapparition, l'idée d'un raisonnement qui ne ferait que rationaliser les conclusions obtenues par d'autres moyens avait disparue. Ainsi, en 1996, quand Evans et Over publient le livre « Rationalité et Raisonnement » ils y préconisent une «théorie du double processus de la pensée» mais avec des processus de type 1 considérés comme rationnels malgré tout (au sens de la rationalité personnelle) et des processus de type 2 mis à jour, passant d'un rôle de pure rationalisation à la base de nos performances logique.

En outre, l'hypothèse de départ que les deux types de processus se produisent en deux temps, d'abord la décision spontanée puis la rationalisation, fut définitivement abandonnée en faveur d'une alternative qui avait été suggéré en 1976, à savoir que les deux types de processus interagissent. Alors que la version antérieure d'Evans et Wason amoindrissait l'idée de rationalité humaine, la version ultérieure d'Evans et Over défend et même étend cette idée.

En 1996 également, le psychologue américain Steven Sloman publie « *The empirical case for two systems of reasoning* » dans lequel il propose une théorie à systèmes duels différente. En 1999, le psychologue canadien Keith Stanovich, dans son livre « *Who is rationnal?*», a utilisé son expertise sur les différences individuelles dans le raisonnement pour proposer une autre approche à système duel. En 2002 Daniel Kahneman reçoit le prix Nobel d'économie avec une autre version à processus duels, dont ses travaux avec Amos Tversky contenaient déjà les prémisses. D'autres chercheurs ont également contribué à cette approche, soit par leur propre version, soit par des critiques de cette approche (pour les critiques les plus fortes voir par exemple Gigerenzer & Gaissmaier 2011; Keren and Schul 2009; Kruglanski, Chenek 2007; Osman, 2004).

Très intuitivement ce que ces théories essaient de capturer est le fait que les humains sont à la fois capables d'avoir des intuitions spontanées et également de penser de manière consciente et délibérée.

Malgré leur diversité ces théories ont en commun la recherche d'une explication aux violations de normativités massivement observées au cours du siècle dernier. Si les décisions des gens ne respectent pas les théories normatives c'est parce qu'elles seraient

basées sur des processus moins coûteux que le raisonnement. Les processus de type 1 sont globalement reconnus comme performants dans la grande majorité des situations « standards » (fréquentes). Dans des cas non-standards (nouveau), en revanche, ils nous font commettre des erreurs.

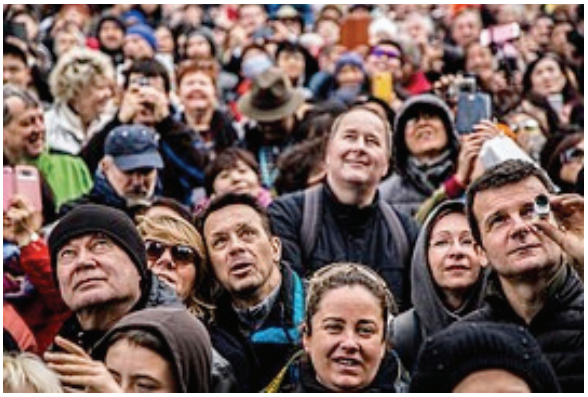
Dans la vision à processus duels, c'est la capacité à laisser le système 2, plus coûteux, prendre le relais du système 1 qui distinguerait les bons des mauvais raisonneurs. Ces théories s'inscrivent parfaitement dans la vision intellectualiste du raisonnement car le système 2 - le raisonnement a proprement parler - permet aux individus de prendre de meilleures décisions en réfléchissant plus longtemps par eux-mêmes. Contrairement aux visions logicistes, cependant, le raisonnement ne réalise pas cela en produisant des inférences valides mais en permettant de surpasser nos intuitions pour limiter nos erreurs intuitives, parfois en allant à l'encontre de théories normatives.

L'enjeu explicatif des théories à processus duels se trouve à la fois dans la caractérisation précise du système 2 et dans la description algorithmique de la façon dont le raisonnement nous permet, parfois, de rester dans le cadre de théories normatives.

Remarquons tout d'abord que, plus largement, en sciences cognitives l'idée de séparer deux type de processus s'est avérée fructueuse dans des domaines aussi divers que l'apprentissage (Reber 1993), la mémoire (Tulving 1985), l'attention (Posner & Snyder 1975), les émotions (Teasdale 1999) ou encore la cognition sociale (Chaiken et Trope 1999).

Prenons simplement deux exemples pour illustrer la diversité de ces approches partageant l'étiquette de « théorie à processus duels » :

Dans le cas de l'attention visuelle, la dichotomie se fait au niveau de l'utilisation d'un même mécanisme de façon « top-down » - contrôlée - ou « bottom-up » - automatique-. Les mécanismes qui vous font reconnaître, de façon automatique, le visage d'un ami passant par hasard dans la rue peuvent également être recrutés pour chercher le visage de votre ami dans une foule de visages. Lorsque vous guidez votre regard dans la foule, testant des zones les unes après les autres, les propriétés du visage de votre ami sont activées en mémoire. Ainsi, lorsque votre regard se posera dans la zone où se trouve le visage de votre ami, vous détecterez 'automatiquement' sa présence.



Chercher le visage d'un ami dans une foule est une utilisation « contrôlée » d'un mécanisme généralement utilisé de façon « automatique ».

Par contraste, pour ce qui est des théories à processus duels dans le domaine de l'apprentissage, la dichotomie est de nature différente. Pour citer un auteur ayant directement inspiré la théorie d'Evans et Over (Evans 2002) : la séparation proposée par Arthur Reber entre apprentissage implicite et explicite n'est pas entendue comme deux utilisations du même mécanisme mais plutôt comme deux systèmes différents. Capter implicitement des régularités de notre environnement d'une côté et écouter la réponse d'un professeur de l'autre nous permettent d'apprendre dans les deux cas, mais cela repose sur des mécanismes cognitifs différents. Reber propose également un lien entre l'accès conscient et les fonctions évolutives d'un système : « on s'attend à trouver que, plus une fonction est montrée primitive, plus elle sera réfractaire à la conscience ». Ce qui correspond assez bien à l'idée que, parmi les deux rationalités invoquées par la plupart des théories à processus duels, l'une serait plus ancienne que l'autre (voir par exemple Stanovich).

Les théories à doubles processus dans le domaine du raisonnement font-elles référence à deux utilisations d'un même mécanisme ? Ou à deux systèmes plus ou moins séparés algorithmiquement ? Cela dépend des auteurs, mais c'est plutôt la deuxième option qui est préférée. Cependant, notons que, parmi les nombreux critères proposés pour séparer les deux systèmes du raisonnement, le contrôle cognitif va être appelé à jouer un rôle majeur.

Nous ne retracerons pas ici l'évolution des différentes théories à processus duels du raisonnement, cela représenterait un tâche bien trop ardue. D'une part à cause de la diversité de ces théories. Pour ne prendre que deux exemples : Sloman fait une séparation entre systéme assossiationiste et système basé sur des règles, alors qu'Evans et Over ou Stanovich considérent que le deuxième système contient plus de choses que l'application de règle (comme la pensée hypothétique ou la capacité à résoudre de nouveaux problèmes). D'autre part, y compris chez ces mêmes auteurs, la caractérisation des deux systèmes n'a cessé d'évoluer, (voir par exemple Evans 2002 p.988, ou Evans 2011 chapitre 1).

Si l'évolution de toute théorie scientifique est essentielle, les théories du raisonnement à processus duels ont été accusées par leurs détracteurs d'être des cibles mouvantes, tant leurs propositions théoriques resteraient vagues et donc pas vraiment falsifiables. Face à ces critiques, on ne peut que saluer l'attitude de chercheurs comme Evans notamment, qui, dans les dernières années a cherché à clarifier ses propositions théoriques, discutant les principaux malentendus du débat et essayant d'écarter certaines critiques récurrentes basées sur ces malentendus (e.g Evans, Chapitre 8 dans Holyoak & Morrison 2012).

De façon encore plus appréciable pour nos objectifs, deux des plus grands noms de la psychologie du raisonnement, Evans et Stanovich, ont tout récemment proposé une théorie commune du raisonnement (Evans & Stanovich 2013). Nous allons maintenant présenter cette théorie qui nous servira dans la suite de cible fixe pour une comparaison avec la théorie argumentative du raisonnement. Nous discutons briévement les principaux problèmes identifiés par les auteurs ainsi que leurs réponses aux critiques qui leur sont le plus souvent addressées.

Le principal objectif de cette thèse, rappelons le, est de proposer un alternative à cette théorie à double processus du raisonnement. Il est donc essentiel, pour nous, de chercher à éviter tout argument de type «straw-man » ou toute simplification abusive de la théorie d'Evans et Stanovich.

Avant de présenter la théorie d'Evans et Stanovich, commençons par présenter des clarifications sur le vocabulaire qu'il nous est conseillé d'employer. Nous avons jusqu'à

présent utilisé « système » ou « processus » de façon quasiment interchangeable. Les premiers articles parus font souvent référence à deux systèmes (e.g Stanovich 1999), mais l'expression de double système s'avère être problématique au moins à deux égards.

D'une part l'idée de deux systèmes peut être comprise comme deux types de systèmes cognitifs ou neuronaux différents, alors que, contrairement à Evans et Stanovich, certains auteurs ne vont pas jusqu'à faire cette hypothèse et n'entendent, par double système, rien de plus que deux types de processus (Kahneman, 2011; S. A. Sloman, 1996). D'autre part, parler de deux systèmes véhicule également l'idée qu'il n'y aurait exactement que deux systèmes cognitifs correspondants aux deux types de processus, ce qui ne traduit pas la pensée de la majorité des chercheurs : le système 1, en particulier, est classiquement associé à une multitude de mécanismes intuitifs que l'on peut nommer plus justement TASS (*The Autonomous Set of Systems,* Stanovich 2009).

Nous utiliserons dans la suite les termes de système 1 et système 2 pour faire référence au(x) mécanisme(s) cognitif(s) responsable(s) de processus de type 1 et type 2. Enfin, nous appellerons « incongruente » une tâche où l'intuition initiale des sujets et une théorie normative pointent vers deux réponses différentes. Une tâche congruente, à l'inverse, permet généralement aux sujets de donner la réponse correcte en suivant leurs intuitions.

3.2 La théorie commune d'Evans et Stanovich

Le premier défi pour une théorie à processus duel est d'offrir une distinction précise entre ce qui relève du processus de type 1 et ce qui relève du processus de type 2.

Un tableau des plus typiques, que l'on retrouve dans la majorité des articles proposant une approche à processus duel du raisonnement, est un tableau contrastant les caractéristiques de chaque type de processus comme le suivant (adapté de Evans et Stanovich 2013) :

Processus de **type 1**	Processus de **type 2**
Rapide	Lent
Grandes capacités	Capacités limitées
Parallèle	Sériel
Inconscient	Conscient
Réponses biaisées	Réponses normatives
Contextualisé	Abstrait
Automatique	Contrôlé
Associatif	Basé sur des règles
Prise de décision basée sur les expériences passées	Prise de décision basée sur les conséquences
Indépendant des capacités cognitives	Corrélé aux capacités cognitives

Evans et Stanovich soulèvent un premier malentendu par rapport à ce genre de tableau : certains chercheurs, critiques vis à vis des théories à processus duels, l'interprètent comme une liste de caractéristiques définissant chacun des deux types de processus. Par exemple, (Krugranslki & Gigerenzer 2011), (Osman 2004) ou encore (Keren et Schul 2009) adressent des critiques qui consistent à mettre en avant des cas où au moins deux traits d'un même type de processus ne sont pas observés en même temps. Pour Evans et Stanovich c'est un argument fallacieux car cette liste de caractéristiques n'est qu'une liste de « corrélations typiques » et non une liste de traits définissant chaque type de processus. En conséquence, le fait d'observer un non-alignement de ces caractéristiques ne constitue en aucun cas un argument contre les théories à processus duels. Pour prendre un exemple trivial, le fait que des processus ayant des caractéristiques de type 1 puissent mener à une réponse normative n'est pas une preuve que la dichotomie est erronée. Les auteurs notent

d'ailleurs que c'est une des critiques fallacieuses les plus fréquentes que de considérer les processus de type 1 comme systématiquement non-normatifs et de type 2 comme systématiquement normatifs.

Certains traits de la liste ont également été repensés de façon relativement nouvelle par rapport au courant logiciste. La pensée de type 2 n'est, par exemple, plus considérée par les théories récentes comme abstraites et sans contexte. Pour ces théories, même si les processus de type 2 sont souvent nécessaires pour résoudre les tâches de raisonnement abstraites - comme celles typiquement étudiées en laboratoire-, cela ne fait pas de la dé-contextualisation une caractéristique qui définirait les processus de type 2.

A ce stade on peut se dire que, si les théories à processus duels ne sont qu'une liste de traits contrastants pas forcément corrélés entre eux, elles apparaissent effectivement comme difficilement falsifiables et pour le moins mal définies. Nous allons voir cependant que la théorie d'Evans et Stanovich est plus précise que cela. Elle propose non seulement deux critères nécessaires et suffisants pour qu'un processus soit de type 2, mais également une proposition algorithmique relativement précise des mécanismes générant ces processus - le système 2-. Leur théorie propose même des considérations évolutionnistes qui auraient pu faire émerger ce système 2 – le raisonnement à proprement parler-.

Commençons par un exemple qui nous semble assez bien synthétiser la vision qu'Evans et Stanovich ont du raisonnement.

Vous êtes un chasseur-cueilleur qui doit prendre une décision importante, nouvelle et difficile. Imaginez par exemple avoir à évaluer les risques que votre enfant vienne à la chasse avec vous. Les mécanismes intuitifs - qu'avaient déjà vos ancêtres – tirant parti d'expériences similaires passées ne vous sont pas d'une grande utilité (disons que c'est votre premier enfant), sauf vous donner l'intuition qu'amener votre enfant serait une bonne idée – imaginons, par exemple, qu'aucun de vos camarades de chasse habituels ne soit disponible-. Face à ce type de situation importante pour vos gènes, nouvelle et contre-intuitive, le système 2 aurait évolué pour vous permettre de faire la chose suivante :

Vous vous imaginez partir à la chasse avec votre enfant, et réalisez au cours de cette simulation mentale que tous les endroits où il serait intéressant de chasser contiennent une

végétation très haute, quasiment aussi haute que la taille de votre enfant. Vous vous ravisez et décidez finalement d'aller à la chasse tout seul pour cette fois.

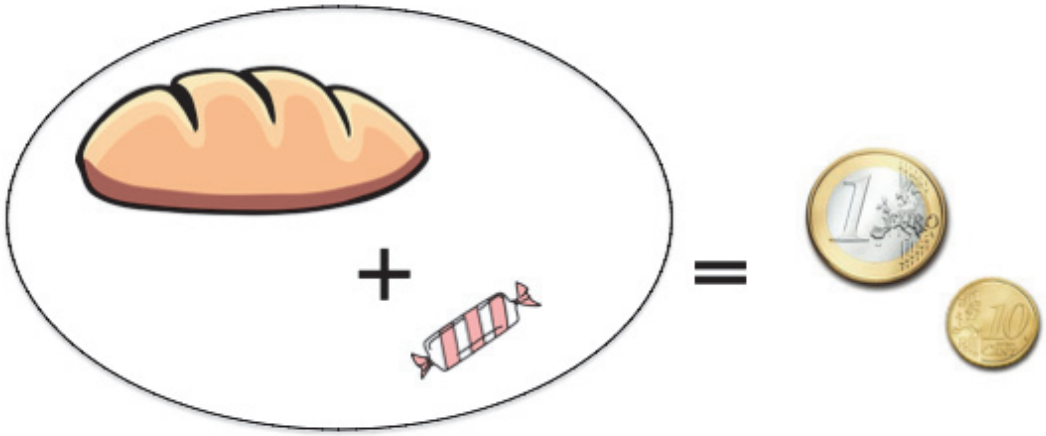
Quasiment tous les éléments essentiels de la théorie d'Evans et Stanovich se trouvent dans cet exemple. Le système 2 est un système cognitif qui a évolué plus tardivement que les mécanismes de système 1. Sa fonction principale est de corriger nos intuitions lorsqu'elle sont mauvaises : amenez votre jeune enfant à la chasse aurait été dangereux. Comment le raisonnement fait-il cela ? Grâce à une capacité de découplage cognitif et à de la simulation mentale. Ces capacités vous permettent de « tester » mentalement des hypothèses en recrutant de façon contrôlée d'autres mécanismes. Ceci permet aux individus de prendre de meilleures décisions dans les cas où nos intuitions sont trompeuses. Présentons le tableau tiré de l'article des auteurs et résumant leur théorie à processus duel :

Processus de type 1	Processus de type 2
Traits définissant	Traits définissant
Ne nécessite pas la mémoire de travail Autonome	Nécessite la mémoire de travail Découplage cognitif, simulation mentale
Traits typiquement corrélés	Traits typiquement corrélés
Rapide, Grandes capacités, Parallèle, Inconscient, Réponses biaisées, Contextualisé, Automatique, Associatif, Prise de décision basée sur les expériences passées, Indépendant des capacités cognitives	Lent, Capacités limitées, Sériel, Conscient, Réponses normatives, Abstrait, Contrôlé, Basé sur des règles, Prise de décision basée sur les conséquences, Corrélé aux capacités cognitives

Système 1 (vieux)	Système 2 (récent)
Apparu tôt dans l'évolution	Apparu tard dans l'évolution
Similaire à la cognition animale	Distinctement humain
Connaissance implicite	Connaissance explicite

Prenons à présent un exemple de problème typique du courant processus-duels. Il nous permettra d'illustrer les critères d'activation du système 2 dans la théorie d'Evans et Stanovich.

Le problème dit du « bonbon et de la baguette » illustré ci dessous, fut présenté pour la première fois par Frederick (2005) dans une série de problèmes ayant une intuition trompeuse, le « *Cognitive Reflection Test* ».



Un bonbon et une baguette coûtent en tout 1.10€.
La baguette vaut 1€ de plus que le bonbon.
Combien vaut le bonbon ?

La majorité des sujets donnent la réponse 0.10€, pourtant la bonne réponse est 0.05€. L'interprétation classique est que les sujets qui répondent 0.10€ interprèteny la deuxième phrase de l'énoncé comme « la baguette vaut 1€ ».

A partir de cette observations d'échecs massifs à un problème simple (au sens ou la réponse est accessible), on peut alors se poser la question suivante : pourquoi notre Système 2 ne vient-il pas corriger notre mauvaise intuition dans ce cas là ? On peut imaginer plusieurs explications dans le cadre de la théorie d'Evans et Stanovich :

Premièrement, et c'est l'explication la plus classiquement invoquée, cela vient du manque de ressources cognitives ou du manque de motivation. Le système 2, par définition, étant plus coûteux à employer que le système 1. L'idée est que notre cerveau ne va pas engager de ressources supplémentaires sans motivation, surtout s'il a déjà une réponse qui semble appropriée. C'est ce que les partisans des théories appellent l'«avarice » ou encore la «paresse » cognitive. Notons qu'Evans et Stanovich proposent trois critères pour l'activation du raisonnement : la motivation, la nouveauté et la difficulté. Dans le cas du bonbon et de la baguette, en plus de la paresse générale du raisonnement, la

structure additive (1.10 = 1 + 0.10) est si entraînée qu'elle amène intuitivement à la réponse fausse. On peut imaginer qu'en un sens cela rend la situation, ni vraiment nouvelle, ni vraiment difficile – ou en tout cas apparaissant comme telle-.

Un deuxième type d'explication qui peut être invoqué provient des travaux de Stanovich sur les différences individuelles. Dans sa théorie commune avec Evans, précisons tout d'abord que le système 2 est en fait divisé en deux sous-systèmes. Il y aurait, d'une part, le « système algorithmique » qui correspond à ce que l'on a décrit jusqu'à maintenant comme le système 2, d'autre part le « système réflexif », qui serait un niveau de contrôle du système algorithmique. Il correspondrait à notre « disposition à la pensée rationnelle ». Ce dernier système influencerait donc le déclenchement du système algorithmique en fonction de notre mode de pensée. Il existe, en outre des différences d'intelligence générale liées à nos capacités de mémoire de travail (correspondant au système algorithmique) : les individus ont, par exemple, plus ou moins tendance à récolter de l'information avant de prendre une décision, ou plus ou moins tendance à penser aux conséquences avant d'agir, etc. Cela fournit un deuxième type de variance pouvant expliquer pourquoi certains individus trouvent la bonne réponse au problème du bonbon et de la baguette et d'autres non.

Notons enfin que Stanovich fait l'observation suivante dans ses travaux sur les différences individuelles : les tests de capacités cognitives ne sont pas de très bons prédicteurs pour savoir quels sujets donnent la bonne réponse à ce genre de problème piège (Stanovich 2010 « *What Intelligent Tests Miss* »). Cette source additionnelle de variance interindividuelle serait, d'après Stanovich, le reflet de différences dans les dispositions à la pensée rationnelle. Ceci pourrait permettre, dans le cadre des théories à processus duels, d'expliquer pourquoi le système 2 échoue relativement souvent à réaliser sa fonction correctrice du système 1 - ce qui, admettons-le, arrive fréquemment en laboratoire comme dans la vie de tous les jours -.

Nous avons désormais une vision suffisamment précise de la théorie d'Evans et Stanovich pour démarrer une discussion critique de celle-ci. Parmi les cinq critiques auxquelles les auteurs se proposent de répondre, nous focaliserons notre attention sur les

deux suivantes. D'une part la réponse qu'Evans et Stanovich donnent aux chercheurs défendant l'idée qu'un seul type de processus peut rendre compte des mêmes observations que leur théorie à processus duels. D'autre part les preuves expérimentales qu'ils vont fournir pour l'approche processus duels, preuves accusées par certains d'être ambiguës et non convaincantes.

La première réponse s'adresse principalement à (Krugklanski & Gigerenzer 2011 et à Gigerenzer 2011) qui prétendent qu'il existe « des arguments et des preuves expérimentales pour une approche théorique unifiée expliquant à la fois les jugements intuitifs et délibératifs comme étant basés sur des règles, par contraste avec l'approche systemes-duels et ses processus qualitativement différents ». La réponse d'Evans et Stanovich a le mérite d'être claire : « this makes no sense to us ».

Pour Evans et Stanovich, le fait que les intuitions et les délibérations soient toutes deux basées sur des règles n'est ni un argument pour, ni un argument contre le fait que ces règles soient basées sur les mêmes systèmes cognitifs. On pourrait, en effet, avoir un seul type de processus produisant ces règles ou bien avoir deux types de processus produisant deux types de règles. Justement, ils considèrent que cet argument est un « straw-man » au sens où , dire que les mécanismes intuitifs (comme avoir peur d'un serpent) sont basés sur des « règles » est pour le moins problématique. Pour les auteurs, Gigerenzer joue sur les mots car, lorsque l'on parle de pensée « basée sur des règles », l'interpréter comme « pouvant être modélisée par des règles » n'est pas vraiment une interprétation charitable. En effet, les « règles » dans le sens d'Evans et Stanovich référent typiquement à l'application de règles logiques ou mathématiques, comme par exemple ce qui vous permet de passer d'une ligne à une autre lorsque vous résolvez une équation mathématiques.

Pour la défense de Gigerenzer, si l'on ne considère pas les règles normatives comme ayant une place particulière dans la pensée humaine, ce qui sépare un mécanisme basé sur des règles d'un mécanisme ayant intégré des régularités dans notre environnement est une distinction loin d'être claire.

Admettons que le fait d'avoir peur quand on voit un serpent n'est pas une règle. Le fait d'éviter une zone de son jardin car il y a des serpents, est-ce une règle pour Evans et

Stanovich ? Probablement pas. Les décisions basées sur les expériences passées sont typiquement de type 1. Dans ce cas là on commence à voir que, si on refuse l'hypothèse que notre esprit intègre et applique des lois logiques, on ne sait plus vraiment à quoi fait référence « un mécanisme basée sur des règles ».

Essayons tout de même de sauver cette idée de règle, même sans hypothèses logicistes. On pourrait penser à des plans d'actions ou à des procédures. Par exemple, si je prépare des œufs au plat, on peut être tenter de dire que j'effectue un plan d'action guidé par des 'règles'. D'abord je fais chauffer la poêle, je mets de la matière grasse, puis je prends les œufs, etc. Même dans ce cas là, cependant, ces procédures peuvent difficilement être considérées comme de type 1 : elles sont clairement tirées d'expériences passées, automatiques et contextualisées. De façon similaire lorsque l'on demande à des lycéens, le jour du bac, de dessiner le tableau de variation d'une fonction, une procédure vient intuitivement en tête (pour ceux qui ont révisé en tout cas !) : calculer la dérivée, faire son tableau de signe, etc. La seule différence avec la préparation d'un œuf au plat c'est qu'ici l'exécution du plan nécessite l'application de règles sur des objets abstraits (pour dériver la fonction par exemple), qui est le seul type de règles auquel Evans et Stanovich semblent se référer lorsqu'ils parlent de « mécanisme basé sur des règles ». C'est en effet le seul type de règles qui pourrait correspondre à des processus de type 2, nécessitant forcément la mémoire de travail. Or, étant donné que, de notre point de vue, il n'y a pas de raison de donner un statut particulier à ces règles, ce que désigne un « mécanisme basé sur des règles » n'a pas de sens clair. Passons à présent aux données expérimentales invoquées par les auteurs comme des arguments forts pour l'approche processus-duels.

Le premier type d'argument expérimental en faveur de la distinction entre les deux types de processus repose sur les études forçant les sujets à répondre rapidement, ou en chargeant leur mémoire de travail, en leur faisant réaliser une tâche cognitive parallèle. Dans ces conditions, lorsque les sujets résolvent une tâche incongruente : d'une part le taux de réussite diminue, d'autre part les erreurs ne sont pas aléatoires mais correspondent bien aux erreurs associées au système de type 1. Par exemple lorsque l'on donne la tâche de Wason en demandant au sujet de répondre le plus vite possible (Robert & Newton 2001) ou en chargeant leur mémoire de travail (De Neys 2006), le taux de bonne réponse est encore

plus bas et les réponses sont guidées par la maximisation de pertinence (menant à la mauvaise réponse dans la tâche originale). Dans le même esprit, les temps de réponses associés aux réponses normatives sont généralement plus longs que les temps de réponses associés aux réponses intuitives.

De notre point de vue, ces données peuvent difficilement être considérées comme des preuves expérimentales d'une distinction entre deux types de processus. La seule chose que ces données montrent est que certains mécanismes sont plus rapides à produire une réponse que d'autres. Leur nature en revanche pourrait très bien être la même.

Dans l'exemple du bonbon et de la baguette par exemple, l'intuition erronée « 0.10€ » vient très clairement des processus de type 1. Personnellement, lorsque le problème m'a été posé pour la première fois, c'est la première réponse que j'ai donnée. Cependant, d'autres intuitions me sont venues en tête: « cela semble trop simple » et « je vais poser l'équation ». C'est en voulant poser l'équation que j'ai vu mon erreur de lecture et l'ai corrigée. J'aurais donc fait partie des sujets qui auraient donné la réponse intuitive avec un temps limité ou en situation en surcharge cognitive, mais qui auraient trouvé la bonne réponse en condition normale. Voilà donc ici un exemple qui semble illustrer parfaitement l'argument d'Evans et Stanovich. C'est pourtant loin d'être évident.

Si on reprend l'exemple des œufs au plat, imaginons que j'ai l'intuition automatique de vouloir manger des sucreries, mais décide finalement d'appliquer plutôt la procédure œuf au plat.

Y-a-t-il une différence de nature entre, d'un côté les processus qui m'auraient amené à manger des sucreries, et de l'autre les mécanismes qui m'aurait fait cuire des œufs au plat ? Non. Que ce soit au moment où je pense à appliquer la procédure œuf au plat, ou durant sa réalisation, aucun processus de type 2 n'est nécessaire. Encore une fois la seule différence entre cet exemple et le problème du bonbon et de la baguette, est que, pour l'application de ma procédure « équation » j'ai effectivement besoin d'utiliser des règles normatives, et donc quasiment, par définition, des processus de type 2. Si d'un côté l'exécution d'un plan d'action comme les œufs au plat se réalise sans soucis, même avec sous charge cognitive, c'est en revanche plus difficile pour poser une équation.

En résumé, les différences temporelles invoquées comme argument pour la distinction des deux types de processus peuvent tout simplement être dues au temps qu'il faut pour appliquer une procédure. Le point important ici est que l'idée d'appliquer des procédures, le choix de procédures, est de type 1, clairement basée sur les régularités des expériences passées.

Ainsi, si l'on ne fait pas l'hypothèse initiale que les mécanismes employés pour appliquer des règles normatives sont de natures différentes que ceux pour employer des règles intuitives, on ne peut pas conclure que les mécanismes arrivant après les premières intuitions sont de nature différente. Notons bien l'aspect tautologique de la chose. D'abord on définit les processus de type 2 comme nécessitant la mémoire de travail. Puis, dans un type de tâches où l'application de procédure menant à la bonne réponse nécessite forcément la mémoire de travail, on montre que réduire les capacités de mémoire de travail laisse les gens avec leur première réponse intuitive. A part si l'on accorde un statut particulier à l'application de procédure normative, l'argument n'est pas convaincant.

Le deuxième type d'argument invoqué pour justifier la distinction entre les deux types de processus consiste en des données neuroscientifiques. Cette partie de l'article est de très loin la partie la moins bien argumentée dans la réflexion des auteurs. Au mieux, les données d'imagerie cérébrale invoquées ne sont pas incohérentes avec les théories à processus duel. Pour ne prendre qu'un exemple, observer que des aires cérébrales différentes s'activent dans le cas d'une réponse normative ou d'une réponse biaisée n'est absolument d'aucun intérêt pour notre discussion. Passons plutôt au deuxième problème majeur pour les théories à processus duel : après la caractérisation des processus nous allons maintenant nous concentrer sur l'interaction entre les deux systèmes.

3.3 Pourquoi le raisonnement ne corrige-t-il pas nos mauvaises intuitions ?

Pour les théories à processus duels, les processus de type 2 ont comme fonction de prendre le relais de nos mécanismes intuitifs lorsqu'ils nous trompent. En particulier, dans la théorie d'Evans et Stanovich, ces processus représentent même un système cognitif

différent, défini par l'utilisation de la mémoire de travail et la faculté d'utiliser des mécanismes intuitifs « hors ligne ». Nous avons vu que, pour expliquer que le système 2 ne corrige souvent pas nos intuitions dans des problèmes incongruents comme celui du bonbon et la baguette, les deux explications principales de ces auteurs sont le manque de ressources cognitives – les processus de type 2 étant coûteux en mémoire de travail – et les différences interindividuelles en termes de dispositions à la pensée rationnelle.

On peut à présent essayer de compléter ces explications au niveau algorithmique, en particulier par des éléments sur l'interaction des deux systèmes. Dans le problème du bonbon et de la baguette, si la majorité des sujets donnent la réponse intuitive, est-ce parce que leur système 2 n'a pas détecté qu'il y avait une erreur normative ? Ou l'aurait-il détecté mais ne serait pas parvenu à prendre le relais ?

Concernant l'interaction entre les deux systèmes, on peut commencer par séparer les théories à processus duel en deux catégories : soit les deux systèmes s'activent de façon parallèle (par exemple Epstein, 1994; Sloman, 1996 ; Ball et Stuppler) soit ils s'activent de manière sérielle (par exemple Kahneman & Frederick, 2005; Stanovich & West, 2000). Les deux cas, nous allons le voir, s'avèrent problématiques pour des raisons différentes.

Le problème de l'activation sérielle est qu'il faut expliquer comment le système 2 décide de prendre le relais. On peut penser qu'il s'active en cas de conflit entre le produit du système 1 et des lois normatives. Mais comment peut-il détecter un tel conflit par rapport à une réponse normative alors qu'il n'est pas encore activé ? Que les processus intuitifs puissent mener à de bonnes réponses dans des cas standards est une chose, mais qu'ils puissent détecter des erreurs normatives ne semble, à priori, pas correspondre à la vision d'Evans et Stanovich.

La version des processus en parallèle est sans doute encore plus problématique que la version sérielle. En effet, si l'on considère que le deuxième système est plus coûteux à employer que le premier, imaginer que ce deuxième système soit activé en permanence serait pour le moins surprenant en terme de rapport coûts / bénéfices.

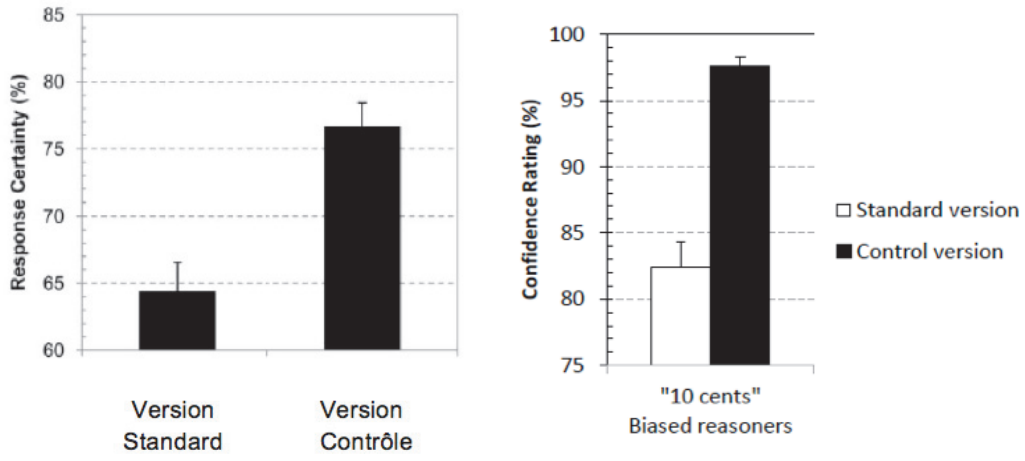
La version sérielle est certes capable d'expliquer certaines données que nous ne discuterons pas ici (voir par exemple Stupple & ball 2008 pour le biais de croyances dans le raisonnement syllogistique). Il est intéressant de remarquer que Edward Stupple et Linden Ball ont, depuis, reconnu le manque de plausibilité de la proposition sérielle et changé de

position (Stupple et Ball, Chapitre 7 Macchi, Bagassi, Viale 2016). Il semble que ce qui les a fait changer d'avis soient des données expérimentales récentes et particulièrement pertinentes dans ce débat : la mise en évidence de ce qui semble être des intuitions normatives, y compris chez les sujets qui donnent la réponse intuitive et donc incorrecte dans des problèmes comme le bonbon et la baguette.

De Neys et collaborateurs ont étudié le sentiment de confiance des participants lorsque qu'ils résolvent des problèmes comme celui de la négligence du taux de base (DeNeys, Cromheeker et Osman 2011) ou encore le bonbon et la baguette (De Neys, Rossi et Houdé 2013), à la fois dans la version originale (incongruente) et dans une version congruente de ces problèmes. Le tableau ci-dessous donne un exemple du matériel expérimental utilisé, les différences entres les deux versions sont soulignés.

	VERSION STANDARD (INCONGRUENTE)	VERSION CONTRÔLE (CONGRUENTE)
Négligence du taux de base	Un psychologue a fait une brève description pour tous individus dans un groupe de 1000 personnes composé de 995 femmes et 5 hommes. La description ci-dessous a été choisie au hasard parmi les 1000 possibles : Camille a 23 ans et est en passe d'obtenir son diplôme d'ingénieur. Le Vendredi soir, Camille aime traîner avec ses amis en écoutant de la musique électro et boire des bières. Laquelle de ces deux affirmations est la plus probable ? a. Camille est un homme b. Camille est une femme	Un psychologue a fait une brève description pour tous individus dans un groupe de 1000 personnes composé de 995 hommes et 5 femmes. La description ci-dessous a été choisie au hasard parmi les 1000 possibles : Camille a 23 ans et est en passe d'obtenir son diplôme d'ingénieur. Le Vendredi soir, Camille aime traîner avec ses amis en écoutant de la musique électro et boire des bières. Laquelle de ces deux affirmations est la plus probable ? a. Camille est un homme b. Camille est une femme
Le bonbon et la baguette	Un bonbon et une baguette coûtent en tout 1.10€. La baguette coûte 1€ de plus que le bonbon. Combien coûte le bonbon ?	Un bonbon et une baguette coûtent en tout 1.10€. La baguette coûte 1€. Combien coûte le bonbon ?

Les même sujets avaient à résoudre les deux formes de problèmes (dans différents ordres et habillages, tout étant contrebalancé). L'observation la plus intéressante est la suivante : dans les versions incongruentes, les sujets qui donnent la réponse intuitive – et donc incorrecte - indiquent une confiance en leur réponse inférieure à la confiance qu'ils affichent dans la version congruente, lorsqu'ils donnent aussi la réponse intuitive – et donc correcte cette fois. Observons la différence de confiance des « mauvais raisonneurs » lorsqu'ils donnent la réponse intuitive à la version congruente et lorsqu'ils donnent la réponse intuitive dans la version incongruente. Les graphiques ci-dessous sont adaptés de (De Neys et al 2011) et (De Neys et al 2013), respectivement.



Dans le cadre des théories à processus duels, ces résultats impliquent que la divergence de processus cognitifs entre d'un côté, les sujets qui donnent la réponse normative et de l'autre ceux qui donnent la réponse intuitive, est tardive. Tous les sujets semblent au départ détecter inconsciemment « qu'il y a un problème ». Pour l'exprimer dans sa version la plus neutre par rapport aux théories normatives, ce sentiment peut être expliqué par des intuitions métacognitives de bas-niveau comme la fluidité « fluency » (Reber et Schwarz, 1999), ou plus généralement un sentiment de certitude « feeling of rightness » (Koriat, 1993).

De Neys et Bonnefon (2013) discutent des implications de ces résultats pour les théories à processus duels et proposent l'idée que ces données contraignent les explications que l'on peut apporter à la question : pourquoi, lorsque les sujets résolvent des problèmes incongruents, certains donnent la bonne réponse et d'autre non ?

Les auteurs classifient les explications généralement apportées à cette question en trois grandes catégories. La raison pour laquelle les sujets qui ne donnent pas la bonne réponse intuitive pourrait être un échec de stockage en mémoire des lois normatives, un échec de détection d'une erreur normative ou bien un échec d'inhibition de la réponse intuitive.

Une explication, en terme de stockage, correspond à l'idée directement tirée du courant logiciste que notre esprit utiliserait des règles formelles. Or, certains sujets pourraient tout simplement ne pas avoir, dans leur répertoire de règles normatives, celle correspondant à la réponse correcte dans le problème donnée. Au-delà des problèmes

théoriques que posent le courant logiciste, cela ne peut pas expliquer les cas comme celui du bonbon et de la baguette ou la tâche disjonctive de Paul et Linda car on peut penser que l'application de la règle normative, dans ces cas là, est parfaitement accessible aux sujets. Nous présenterons par la suite des données montrant que la majorité des sujets acceptent quasiment instantanément la bonne réponse lorsqu'elle leur est expliquée. Les cas que pourrait éventuellement expliquer cette idée d'échec de stockage seraient ceux des jeunes enfants, qui, dans le cas du bonbon et de la baguette par exemple, auraient des difficultés pour ne serait-ce que comprendre l'expression « de plus que », expression linguistique posant particulièrement problème au début de la scolarité (Vergnaud).

Comme De Neys et Bonnefon le soulignent, très peu de chercheurs contemporains défendent cette hypothèse si ce n'est pour l'explication de biais particuliers. Les implications les plus intéressantes à tirer à partir des observations dites « d'intuitions logiques » se trouvent plutôt entre les explications en terme d'échec de contrôle et en termes d'échec d'inhibition.

L'explication en terme de contrôle serait sans doute celle que fournirait la théorie Evans et Stanovich. Elle consiste à dire que les sujets qui donnent la réponse intuitive dans des problèmes incongruents ne détectent pas le conflit existant entre leur intuition et la réponse normative. Les sujets auraient les moyens d'arriver à la réponse normative mais leur esprit ne se rendrait jamais compte qu'il y a une erreur. Dans cette hypothèse, la différence avec les sujets qui donnent la bonne réponse se situe très tôt dans le déroulement de la résolution du problème. Contrairement aux « mauvais raisonneurs », le cerveau des « bons raisonneurs » détecterait de façon implicite que leur première intuition viole une norme, ce qui déclencherait l'activation du système 2 et les amènerait à la bonne réponse.

Les données de De Neys et collaborateurs vont clairement à l'encontre de cette explication en montrant que même les sujets donnant la réponse intuitive à un problème incongruent détectent qu'il y a un problème avec leur première intuition. La différence avec les « bons raisonneurs » se trouverait donc plutôt dans la capacité à utiliser ce signal de conflit pour inhiber leur première intuition. Il pourrait alors exister plusieurs explications pour cet échec d'inhibition. Cela pourrait être dû à un manque de ressources cognitives

et/ou de motivation, ou au fait qu'une réponse intuitive ne pourrait être inhibée que si le signal de conflit est suivi par des processus délibératifs de type 2 fournissant des justifications explicites pour remettre en cause la première intuition. Sans ce type de validation explicite, les sujets pourraient ne pas changer leur première réponse intuitive.

A quel point ces récentes expériences sont-elle un problème pour la théorie d'Evans et Stanovich ? A priori l'idée même d'intuition logique semble paradoxale pour une théorie qui prend comme base une séparation entre intuition et pensée normative. Cependant, étant donné que le fait d'être sensible aux théories normatives n'est pas une propriété définissant le système 2 mais simplement une corrélation typique, des intuitions peuvent, en théorie, s'avérer être normatives. Originellement, on peut penser qu'Evans et Stanovich ont pris ces précautions théoriques pour rendre compte des cas congruents où intuitions et réponses normatives sont alignées. Cependant, lorsqu'ils discutent de ce qui rend un environnement favorable ou hostile au système 1, ils mentionnent la présence d'indices dans l'environnement (ici l'énoncé du problème) que les sujets pourraient exploiter grâce à leurs expériences passées et leurs systèmes 1. Comme De Neys et Bonnefon l'admettent eux-mêmes, des éléments saillants dans l'énoncé pourraient jouer un rôle dans la détection de conflit. Par exemple, dans le cas de problèmes de négligence du taux de base exhibant un taux extrême, on peut penser que cela pourrait être un indice exploitable par le système 1.

En revanche, dans des problèmes comme le bonbon et la baguette, on voit a priori mal comment le système 1 pourrait générer un signal d'erreur à partir de l'environnement et de la pratique. En effet, ce qui amène les sujets à l'erreur intuitive semble précisément être le surentraînement des sujets à résoudre des problèmes additifs simples. Le fait que les sujets détectent cette subtile erreur d'interprétation pourrait cependant venir d'un signal d'incohérence interne venant de mécanismes de bas-niveau. On peut penser par exemple que les mécanismes responsables de la lecture fournissent un signal que « tout ne s'est pas passé de manière parfaitement fluide ». Entre la version congruente et incongruente du bonbon et de la baguette, il y a au moins une phrase de l'énoncé incongruent qui n'aurait pas été traitée de manière aussi fluide que dans la version congruente. Evans et Stanovich pourraient alors essayer d'étendre leur idée d'indice dans l'environnement que le système 1

pourrait exploiter. Ils pourraient par exemple défendre que dans le cas du bonbon et de la baguette, ce serait l'expérience des mécanismes intuitifs de lecture qui fournirait un signal d'incohérence que les sujets pourraient ensuite exploiter ou non. Ces processus de type 1 ne seraient donc pas des intuitions normatives, mais un sentiment métacognitif intuitif que « quelque chose ne s'est pas déroulé comme d'habitude » lors de la lecture de l'énoncé.

Quoiqu'il en soit, les résultats montrent que même les sujets qui finissent par donner la mauvaise réponse dans des problèmes incongruents ont un signal, même faible, que quelque chose s'est mal passé. Et ceci devrait forcément pousser la théorie d'Evans et Stanovich à au moins revoir l'importance des différences interindividuelles dans leur théorie.

Ils ne peuvent plus vraiment dire que les sujets qui donnent la réponse intuitive au problème de la négligence du taux de base ou du bonbon et de la baguette ne détectent pas du tout la violation de norme. Il semble qu'ils devront au moins ajouter l'idée qu'il y a une détection inconsciente, ce qui revient à l'hypothèse d'un échec d'inhibition. Evans et Stanovich pourraient ensuite défendre que ce ne sont pas des intuitions normatives mais des intuitions de type 1 sur le fait que « tout ne s'est pas passé comme d'habitude», ce qui, dans le cas de ces tâches incongruentes, serait un bon « proxy » pour le non-respect de normes.

Pour conclure ce chapitre sur les théories à processus duels, notons que De Neys et Bonnefon (2013) discutent également des théories du raisonnement qui rejettent l'idée que la logique ou la théorie des probabilités soient des normes appropriées pour le raisonnement. En particulier les auteurs mentionnent les partisans de normes bayésiennes, de normes logiques non-classiques mais également de normes pragmatiques guidant les échanges conversationnels dont, en particulier, la théorie argumentative du raisonnement. Discutons donc leur propos concernant ces normes alternatives.

La remarque de De Neys et Bonnefon à propos des normes alternatives est la suivante: ces théories sont orthogonales à la distinction entre échecs de stockage, échecs de contrôle, et échecs d'inhibition au sens où ces théories pourraient être, à priori compatibles avec les trois explications.

En effet, pour les théories alternatives, il est probable que les sujets n'aient pas de règles formelles menant à la réponse normative stockée. D'autre part les sujets pourraient très bien générer une réponse valide dans la norme alternative sans jamais détecter le moindre confit avec la réponse valide dans la norme classique. Enfin les sujets pourraient également détecter un conflit entre la norme alternative et la norme classique sans réussir à inhiber la réponse valide dans la norme alternative.

Nous proposerons dans le chapitre 5 de situer la théorie argumentative par rapport à cette discussion, mais formulons dès maintenant un raffinement de cette dernière option. Les sujets pourraient générer une réponse dans la norme alternative non pas en détectant la violation d'une théorie normative mais simplement avec un signal de conflit interne venant d'intuitions métacognitives sur la fluidité. De plus, ce qui distingue ceux qui utilisent ce signal des autres pourrait simplement venir des sujets les plus entrainés aux pratiques culturelles du raisonnement formel, ou simplement à la pratique du concept « d'énigme » et de « piège ». Ils accepteraient alors le « jeu normatif» que l'expérimentateur semble attendre d'eux. Sans forcément invoquer le concept d'inhibition ces sujets renoncent à leur réponse valide dans la norme alternative pour se conformer à ce qu'on semble attendre d'eux. Ils déploient alors des outils formels, plus ou moins internalisés selon l'expérience des sujets. Dans ce cadre, la validité de l'observation d'Evans disant que mettre l'accent sur le respect des règles logiques dans la consigne augmente le taux de réponse logique (Evans 2002) n'aurait rien d'étonnant.

Cependant il reste à expliquer pourquoi les autres sujets s'entêtent dans leur première intuition et bien qu'ayant pour la plupart probablement déjà rencontré ce type de « problème-piège » dans le passé, ne détectent jamais consciemment leurs erreurs. Pire encore, ils semblent sûrs de leurs erreurs, moins confiants que pour un problème congruent mais tout de même à 82%.

Nous allons à présent présenter la théorie argumentative du raisonnement qui se situe donc comme une des théories proposant une norme alternative aux normes classiques pour juger des compétences du raisonnement humain. La théorie argumentative fait également partie des théories qui, comme celle de Gigerenzer et collaborateurs, considère

que la séparation entre deux types de processus est artificielle, voire tout simplement fausse. La façon de rendre compte des performances du raisonnement humain avec un seul type de processus est, en revanche, bien différente des théories de Gigerenzer et collaborateurs. En particulier nous verrons qu'un seul type de mécanismes intuitifs permet, malgré tout, de proposer une différence entre deux types d'inférences.

Le chapitre suivant propose une alternative aux théories à processus duels à un seul type de processus intuitif et, même, un module. Nous allons le voir, l'idée que le raisonnement est un mécanisme intuitif et spécialisé pourra malgré tout nous faire retomber sur l'intuition initiale des théories à double processus : les humains sont à la fois capables d'avoir des intuitions et de penser de manière consciente et délibérée.

Chapitre 4 - La théorie argumentative du raisonnement

L'hypothèse de fond, que les deux courants de la psychologie du raisonnement décrits dans les chapitres précédents ont en commun, c'est l'idée que la fonction du raisonnement est d'aider les raisonneurs solitaires à dépasser leurs intuitions, leur permettant de prendre de meilleures décisions, et d'avoir de meilleures croyances. Nous avons regroupé les recherches qui font cette hypothèse sous le terme de théories « intellectualistes » du raisonnement. Tenant pour acquis le fait que le raisonnement améliore la pensée individuelle, les recherches actuelles du domaine se sont donc logiquement concentrées sur l'explication des échecs du raisonnement individuel. Leur projet scientifique peut se résumer par le fait d'expliquer ce qui a pu mal se passer pour les sujets qui n'ont pas surpassé les biais de leurs mécanismes intuitifs.

L'objet de ce chapitre est de présenter une alternative à ce projet de recherche. Nous proposerons une définition différente des mécanismes du raisonnement au niveau algorithmique et nous présenterons également une autre caractérisation du raisonnement au niveau ultime, évolutionniste.

Si l'on remet en question l'idée que la fonction du raisonnement est d'aider les individus à faire des inférences logiques ou de leur permettre de corriger leurs erreurs intuitives, peut-être y a-t-il d'autres éléments à apporter, pour expliquer les échecs individuels observés en laboratoire, que le manque de ressources, la perturbation venant d'autres mécanismes ou le manque de disposition à la pensée rationnelle ? Mais si le raisonnement n'est pas ce mécanisme qui permet aux individus d'atteindre de meilleures croyances et de prendre de meilleures décisions grâce à une réflexion personnelle, quelle pourrait bien être sa fonction ? Dan Sperber et Hugo Mercier ont proposé l'idée que la principale fonction du raisonnement est sociale et, en particulier, qu'elle serait argumentative. La théorie argumentative du raisonnement est articulée en tant que telle dès 2009 dans la thèse d'Hugo Mercier mais les premiers éléments apparaissent déjà dans Sperber (2000, 2001) et l'article de référence sortira en 2011 dans la revue Behavioral & Brain Science sous le titre « *Why do humans reason? Arguments for an argumentative theory* ». L'idée de départ, défendue par les auteurs, est que la fonction du raisonnement humain se comprend mieux dans le cadre de l'évolution de la coopération et surtout de la communication humaine. Commençons par présenter quelques éléments sur ces considérations évolutionnistes à partir desquelles ces auteurs vont construire leurs hypothèses concernant les traits que doit avoir le raisonnement pour réaliser sa fonction. Dans le but de cerner au mieux les prédictions expérimentales faites par la théorie argumentative - ce qui sera

l'objet des chapitres suivants - nous allons essayer de présenter, de la façon la plus précise et la plus intuitive possible, la vision interactionniste du raisonnement de Dan Sperber et Hugo Mercier en nous inspirant largement, dans ce chapitre, de leur livre en préparation « *The Enigma of Reason* » (Sperber & Mercier, in prep).

Pour ces auteurs deux grands problèmes, rencontrés par notre espèce au cours de l'évolution, ont façonné nos mécanismes de raisonnement : la coopération et la communication.

4.1 Le défi de la coordination.

La coopération humaine se distingue non seulement par son ampleur mais également par sa diversité et sa flexibilité. Si la coopération animale laisse peu de place à la créativité, chez l'humain, en revanche, les tâches et les partenaires de coopération sont testés, changés, conservés au fil des opportunités. Si une fourmi peut parfaitement prédire le comportement des membres de sa colonie les humains, eux, doivent constamment se poser la question de ce qu'il peuvent attendre de telle ou telle personne, vérifier les bienfaits des alliances passées, en cours, ou possibles. De plus, même quand tous les membres sont coopératifs, il peut y avoir des malentendus dans les buts communs ou dans le rôle de chacun. Comment les humains parviennent-ils à savoir ce qu'ils peuvent attendre des autres ? Comment parviennent-ils à former des attentes mutuelles ?

Deux réponses sont classiquement invoquées : les normes sociales et la psychologie naïve. Certaines normes précises et spécialisées, comme le code de la route, assurent la coopération de façon à ce qu'elle soit bénéfique pour les agents impliqués. En revanche, les textes de loi par exemple, contribuent certes à la coordination mais ce n'est sans doute pas là leur fonction principale. Dans un mariage un des époux sait ce qu'il est en mesure d'attendre de la part de son partenaire en vertu de la loi ou même de la morale. Mais pour atteindre le niveau de coopération d'un couple marié dans la vie de tous les jours un partenaire doit comprendre bien plus que les normes sociales sanctionnées par l'article de loi sur le mariage. Les normes ne suffisent pas car les interactions sociales, dans la plupart des formes qu'elles prennent, laissent une grande place à la créativité et à l'improvisation.

Prenons un exemple : vous organisez une fête chez vous. Qui allez-vous inviter ? Pour certains le choix peut-être guidé par des conventions sociales comme le fait d'inviter les personnes qui vous ont invité auparavant. Mais pour d'autres, le principe de réciprocité n'aide en rien. Pourtant il y a de vrais problèmes de coordination à résoudre. Par exemple, si vous invitez Audrey, mieux vaut ne pas inviter Romain. Si vous invitez Martina, elle voudra que vous invitiez aussi Manuela.

Sachant que vos futures interactions dépendront de l'ensemble des microdécisions interconnectées que vous allez prendre comment résoudre les problèmes de coordination que les

normes ne résolvent pas ? La réponse classique est la psychologie naïve. Vous devez comprendre l'état d'esprit de Audrey, Romain, Mina et des autres possibles invités, anticiper leur réaction et s'assurer que la fête sera un succès

Tout comme les normes, la psychologie naïve joue un rôle essentiel pour le challenge de la coordination qui renforcera vos relations avec les autres et ne les affectera pas. Mais le tableau est loin d'être complet. En effet, les individus n'infèrent pas ce à quoi ils peuvent s'attendre des autres uniquement sur la base des normes sociales et de la psychologie naïve. Les attentes mutuelles de chacun sont discutées, révisées et négociées en détail. La plupart de nos décisions portant sur nos interactions sont elles-mêmes prises de façon interactive.

Le commérage («*gossip* » en anglais) est une des formes de communication humaine les plus fréquentes. Ce type d'interaction à propos d'interactions nous apporte des informations précieuses sur ce qu'on peut attendre de tel ou tel individu. Les individus sont cependant loin de rester les objets passifs des commérages sur leur compte. Ils peuvent participer à cette discussion en justifiant leurs décisions, leurs points de vue et, ce faisant, ils protègent leur réputation.

En nous justifiant nous faisons en fait plusieurs choses. Nous influençons la façon dont les autres comprennent nos états mentaux, jugent notre comportement et parlent de nous. En invoquant une justification pour un comportement nous encourageons les autres à attendre de notre futur comportement qu'il soit guidé par des raisons similaires et à nous le reprocher si ce n'est pas le cas. Si, par exemple, le jour de votre fête vous vous justifiez auprès d'un ami de ne pas avoir invité Jean car vous le trouvez très ennuyeux, vous n'avez pas seulement justifié l'absence de Jean à votre soirée : votre ami s'attend par exemple à vous voir éviter Jean à l'avenir, il ne sera en tout cas pas surpris si c'est le cas. Si vous organisez, le mois suivant, un week-end avec Jean, votre ami est en droit de vous questionner.

En invoquant une justification vous indiquez également que vous êtes susceptible de juger le comportement des autres sur la base de raisons similaires à celles que vous invoquez pour vous justifier vous-même. Dans cet exemple de la fête, vous pourriez justifier auprès d'un collègue d'équipe, le fait d'avoir invité Pierre « car il fait partie de l'équipe ». Lorsque votre collègue organisera à son tour une soirée, même si cette règle ne le satisfait pas, il s'attend à ce que vous le jugiez négativement s'il ne l'applique pas.

Enfin, fournir une justification vous engage dans une discussion où les autres peuvent aussi bien accepter ou questionner vos justifications et faire eux-mêmes appels à des raisons. De telles conversations aident les individus à se coordonner et peuvent faire progressivement émerger des normes sociales.

Réduire les mécanismes de coordination sociale aux respect des normes, à la psychologie naïve ou à un mélange des deux, c'est ignorer à quel point les interactions humaines ont lieu dans le but de se justifier ou d'évaluer les raisons des autres (celles qu'ils nous donnent et celles que nous leur attribuons), de critiquer les interactions passées ou présentes et d'anticiper celles à venir.

L'utilisation de raisons se trouve, en fait, exactement entre les normes sociales et la psychologie naïve. Quand nous nous justifions, nous présentons nos motivations comme bonnes normativement et nous présentons des normes comme étant des motivations. En d'autres termes, nous « psychologisons » les normes et nous normalisons les états mentaux. En réalisant cela notre but n'est pas de fournir une explication objective, sociologique ou psychologique de nos actions et interactions. Notre but est de nous coordonner de manière avantageuse en protégeant, en augmentant notre réputation et en influençant la réputation des autres.

Notons bien que nous ne parlons pas encore de raisonnement ici mais simplement du rôle des raisons. La fonction d'attribuer des raisons à soi et aux autres se comprend donc dans le cadre du défi de la coordination : elle est de se justifier et d'évaluer les justifications des autres - que ce soit celles que nous leur attribuons ou celles qu'ils nous donnent.

Ce n'est cependant pas la seule fonction de l'attribution de raisons, elle a aussi une autre fonction qui sera celle du raisonnement à proprement parler : l'argumentation. Pour comprendre la deuxième fonction de l'attribution de raison et la fonction principale du raisonnement humain, abordons à présent l'autre grand défi dans l'histoire de l'humanité : tirer des bénéfices de la communication.

4.2 Le défi de la communication

La communication humaine frappe par son ampleur, sa diversité et sa complexité. Si la communication offre de nombreux bénéfices, elle présente aussi un risque majeur, celui de se faire manipuler. En effet, si une partie des enjeux de la communication se trouve dans la compréhension mutuelle, un locuteur ne veut pas seulement être compris mais également être crû. Il veut avoir de l'influence sur son audience. Un récepteur ne veut pas simplement comprendre ce que la source veut dire mais il veut aussi, ce faisant, apprendre quelque chose sur un état du monde.

Pour que toute communication puisse être stable évolutionnairement elle doit bénéficier à la fois au récepteur et à l'émetteur (Dawkins & Krebs). Si les émetteurs arrivent trop facilement à influencer les récepteurs le risque de se faire manipuler devient trop grand pour les récepteurs par rapport aux bénéfices épistémiques qu'ils peuvent tirer de la communication. Il vaudrait mieux pour

les récepteurs, dans ce cas, d'arrêter d'écouter. Si les récepteurs arrêtent d'écouter ou simplement s'ils rejettent trop d'informations, il n'y a plus d'intérêt à émettre et la communication s'effondre.

Pour répondre au problème de l'honnêteté de la communication, les humains s'appuient sur le filtrage de l'information communiquée. Nous regrouperons ces mécanismes sous le terme de « vigilance épistémique » (Sperber et al 2010).

La vigilance épistémique est composée de deux mécanismes principaux, l'un nous aidant à savoir *qui* croire, l'autre nous aidant à savoir *que* croire.

Qui devons-nous croire, quand et sur quel sujet et sur quelle question? Nous suivons des conseils différents selon qu'ils viennent d'un avocat ou d'un dentiste. Nous croyons plus facilement des témoins qui n'ont aucun intérêt personnel en jeu. Nous sommes en alerte quand les gens hésitent ou, au contraire, insistent trop. Nous prenons en compte la réputation d'honnêteté et de compétence des autres (voir par exemple Petty & Wegener 1998). De nombreuses expériences ont montré que les enfants développent, à un âge précoce, la capacité de traiter des indices de compétence et de bienveillance pour décider à quelle source faire confiance (voir Harris 2007 ou Clément 2010 pour des revues de littérature).

La confiance en la source ne fait cependant pas tout. Certains contenus sont plus crédibles que d'autres. Pour prendre des cas extrêmes, vous n'accepterez jamais que 1+1 = 5 ou que vous êtes mort. Si vous recevez ces informations des personnes en qui vous avez le plus confiance, vous les prendrez probablement au second degré. A l'inverse, si la personne en qui vous faites le moins confiance vous dit que 6 est plus grand que 3 ou que vous avez besoin d'eau pour vivre, vous serez d'accord.

D'où qu'elles viennent, il est très peu probable que les absurdités soient prises au sens littéral et que les truismes ne soient pas acceptés. La plupart des informations se trouvent évidemment entre ces deux extrêmes, ni évidemment vraies, ni évidemment fausses. Ce qui rend la plupart des informations plus ou moins crédibles est la façon dont elles correspondent à ce que l'on sait. Si l'on se rend compte que le message est incohérent avec ce que nous croyons déjà, nous sommes susceptibles de le rejeter. Pourtant, rejeter un message remettant en question nos croyances peut nous faire manquer une occasion de les réviser de manière appropriée, parce qu'elles étaient fausses ou n'étaient plus à jour.

La vigilance vers la source et la vigilance vers le contenu peuvent également pointer dans des directions différentes. Si une source en qui nous avons confiance émet une information qui contredit nos croyances, une révision de croyances est inévitable. Si nous acceptons l'information,

nous devons réviser les croyances préalables qu'elle contredit. Si nous rejetons l'information, nous devons revoir notre jugement de confiance en la source.

Bien qu'elle soit loin d'être infaillible la vigilance épistémique nous permet, en tant que public, de trier les informations communiquées en fonction de leur fiabilité et permet, généralement, que la réception d'information nous soit bénéfique. Pour les communicateurs, la vigilance de leur public diminue les avantages qu'ils pourraient attendre s'ils cherchent à tromper les autres, elle réduit leurs chances de succès et, s'ils se font détecter, elle augmente les coûts qu'ils pourraient devoir payer sous la forme d'une perte de crédibilité et de réputation.

Ces précautions ont cependant un prix. Même si la vigilance épistémique est dans l'ensemble bénéfique, cette prudence informationnelle se traduit par des occasions manquées. Par exemple vous ne suivez pas le conseil d'un inconnu qui vous indique un bon coin à fraises ou un site qui permet de gagner de l'argent en ligne. De précieux messages peuvent être rejetés par manque de confiance envers le messager : autant d'occasions manquées.

La vigilance épistémique est nécessaire mais elle crée un goulot d'étranglement dans le flux d'informations. Pourtant, pour les récepteurs de l'information, les avantages d'une vigilance épistémique bien calibrée sont supérieurs aux coûts. Pour les émetteurs, par contre, la vigilance de leur public est coûteuse. Cette vigilance pose clairement des problèmes aux sources malhonnêtes, mais aussi aux sources honnêtes. Un émetteur honnête peut vouloir communiquer une information vraie et pertinente, mais ne pas avoir suffisamment d'autorité aux yeux de son interlocuteur pour que l'information soit acceptée. Dans ce cas ils y perdent tous les deux : l'émetteur en influence, et le récepteur en connaissances pertinentes sur le monde. Plus une information est pertinente plus vous devriez vous en méfiez, mais plus vous en priver peut être dommageable.

C'est là où le raisonnement a un rôle à jouer. L'utilisation argumentative des raisons aide de bonnes informations à traverser le goulot d'étranglement que la vigilance épistémique crée dans le flux social de l'information. Le raisonnement est bénéfique aux destinataires car il leur permet de mieux évaluer les informations potentiellement utiles qu'ils n'accepteraient pas sur la seule base de la confiance. Le raisonnement est également bénéfique pour les communicateurs car il leur permet de convaincre un public méfiant.

En tant qu'émetteurs, nous nous adressons à des personnes qui, si elles ne nous croient pas sur la seule base de la confiance, vérifient le degré de cohérence de ce que nous leur disons avec ce qu'elles croient déjà sur la question. Etant tous à la fois émetteurs et récepteurs, nous sommes en mesure de comprendre comment, quand nous communiquons, notre public évalue ce que nous leur

disons. Nous tirons profit de cette compréhension et adaptons nos messages en conséquence. À moins que nous n'essayions de tromper notre public, nous ne devrions pas voir leur vigilance comme un simple obstacle à nos objectifs de communication. Au contraire, nous pourrions être en mesure d'utiliser cette vigilance d'une manière qui sera bénéfique à la fois pour eux et pour nous.

Une bonne façon de convaincre les récepteurs est de les aider à activement vérifier la cohérence de nos messages avec ce qu'ils croient déjà. Nous pouvons les aider à s'apercevoir que, compte tenu de leurs croyances, il serait moins cohérent pour eux de rejeter nos demandes que de les accepter. En d'autres termes, en tant qu'émetteur s'adressant à un public vigilant, nos chances d'être cru peuvent être augmentées en faisant un affichage honnête de notre cohérence, cohérence que notre public aurait de toute façon vérifiée. Un bon argument consiste précisément dans l'affichage de relations de cohérence que le public pourra lui-même évaluer.

En tant que destinataire, lorsque nous disposons non seulement d'une information, mais aussi d'un argument en sa faveur, nous pouvons évaluer l'argument et, si nous le jugeons bon, finir par accepter à la fois l'argument et l'information. Les arguments de votre interlocuteur peuvent être avantageux pour vous de deux façons : d'abord, en affichant la cohérence que vous cherchez à évaluer de toute façon, il vous est plus facile d'évaluer l'information, ensuite si cette évaluation entraîne votre acceptation d'informations pertinentes, il rend la communication plus bénéfique.

Nous construisons des arguments quand nous essayons de convaincre les autres et nous évaluons les arguments donnés par d'autres comme un moyen imparfait mais utile pour reconnaître les bonnes idées et rejeter les mauvaises. Tantôt communicateurs, tantôt auditeurs, nous bénéficions à la fois de la production d'arguments à présenter aux autres et de l'évaluation des arguments que les autres nous présentent. Le raisonnement implique deux capacités : la production et l'évaluation d'arguments. Ces deux capacités sont mutuellement adaptées et doivent avoir co-évolué. Ensemble elles réalisent la fonction principale du raisonnement : la fonction argumentative.

4.3 Comment fonctionne le raisonnement ?

Nous avons jusque-là décrit deux fonctions - deux histoires phylogénétiques - pour un même mécanisme traitant de raisons. Passons maintenant à une description plus proximale du raisonnement. Comment fonctionne le raisonnement ?

Avant tout, remarquons que les fonctions justificatives et argumentatives sont deux usages du même mécanisme traitant de raisons. Une même représentation peut être utilisée à la fois comme une

raison justifiant une action passée et comme une raison argumentant pour influencer une décision future.

Notons par exemple (a) l'énoncé suivant: "Je suis devenu ami avec Lola".
Seul, c'est une simple assertion. Vous pouvez énoncer (a) simplement pour tenir un ami au courant de l'état du monde.

Elle peut être utilisée à des fins justificatives : vous pouvez par exemple dire (a) à un hôte pour justifier le fait d'avoir invité Lola à votre soirée ou de lui avoir fait confiance.

Elle peut aussi être utilisée à des fins argumentatives : vous pouvez dire (a) à un ami pour le convaincre d'inviter Lola à sa soirée ou de faire confiance à Lola.

Autre exemple : vous pouvez utiliser le fait que "Alain Chabat joue dans le Film Rrrr" comme un argument pour convaincre un groupe d'amis de regarder ce film. Mais vous pouvez aussi l'utiliser comme une justification auprès de vos amis pour leur avoir fait regarder ce film (imaginez qu'ils ne l'aient pas aimé et vous disent ne plus vous faire confiance pour ce genre de choix).

Le degré avec lequel vos amis (et vous-même) aiment Alain Chabat fera de cette raison un bon argument ou une bonne justification.

La différence dans le cas d'une justification est que, même si vos amis n'aiment pas Alain Chabat, si vous aviez de bonnes raisons de le croire au moment de votre conseil et qu'ils ne se sont pas manifestés, votre raison est une justification acceptable. Dans le cas d'un argument, en revanche, vos amis vont s'empresser de répondre à votre argument qu'ils n'aiment pas cet acteur. Si vos amis ne détestent pas non plus l'acteur, vous pourrez toujours tenter d'autres arguments pour votre film, la discussion a toutes les raisons de ne pas être terminée.

Si justification et arguments sont deux usages des raisons, nous ne parlerons désormais que *du* raisonnement comme un unique mécanisme traitant de raisons, dont la fonction est d'influencer les autres et de ne se faire influencer qu'à bon escient. Que le raisonnement soit utilisé pour traiter des justifications ou traiter des arguments nous importe finalement assez peu pour la description algorithmique du mécanisme de raisonnement. Comment la production et l'évaluation de raisons se font-elles au niveau algorithmique ?

Reprenons la conversation précédente, si la majorité de vos amis vous répondent ne pas aimer Alain Chabat, vous devez essayer autre chose pour les convaincre. Il vous faut alors chercher, dans l'espace de vos représentations, des arguments qui pourraient marcher. En production, le raisonnement doit trouver des représentations qui montreraient à votre interlocuteur qu'il est plus

cohérent pour lui de choisir votre film plutôt qu'un autre. C'est là sans doute la plus grande prouesse du raisonnement : chercher des raisons adaptées à la situation dans l'espace de ses représentations et ceci au rythme de la conversation et du *feedback* des interlocuteurs. Le film que vous proposez à vos amis peut être mis à tout moment en concurrence avec un autre film (ou même une autre activité). Selon le concurrent, votre raisonnement va se servir des propriétés qui contrastent avec votre film (dans l'exemple: pas une comédie, long, dont aucun acteur n'est connu, pas français, etc.) et évalue parmi les possibles représentations celles qui pourraient, une fois présentées à l'auditoire, produire le meilleur effet dans la conversation en cours. Produire des raisons adaptées au rythme d'une discussion est une prouesse du quotidien qui demande finalement si peu d'efforts que nous en avons oublié à quel point c'est un exercice bien plus exigeant cognitivement, et nécessitant bien plus de travail computationnel que dériver une fonction ou résoudre le problème du bonbon et de la baguette.

En évaluation, nous avons des intuitions sur des liens de cohérence entre les représentations. Si quelqu'un vous dit les deux énoncés suivants l'un après l'autre :

- Jean s'est cassé la jambe.
- Je l'ai vu avec un plâtre.

Plutôt que de prendre ces deux assertions comme séparées et donnant deux informations sur l'état de Jean, vous tendez à comprendre intuitivement la seconde comme une raison d'accepter la première. Les liens raison-conclusion font partie de notre ontologie du monde, comme certaines émotions des autres ou les couleurs d'une certaine longueur d'ondes. Les raisons sont une des dimensions avec lesquelles nous comprenons le monde.

Les intuitions sur les raisons peuvent être à propos de n'importe quel état du monde car elles capturent des propriétés de nos représentations et non des propriétés du monde directement comme les modules visuels traitant des visages ou de la couleur. Le module du raisonnement est méta-représentationnel, il représente quelque chose à propos de nos représentations. Cela rend le raisonnement à la fois très général car il peut traiter de raisons à propos de n'importe quoi, mais il opère en même temps sur une propriété très spécifique de nos représentations : les liens raison-conclusion. Comme tout mécanisme spécifique il peut être appelé module et son traitement est attendu comme étant intuitif et rapide.

Dans l'exemple du groupe d'amis choisissant un film, de nombreuses évaluations de raisons ont lieu. Vos amis évaluent vos arguments pour ne pas choisir un film qu'ils peuvent ne pas aimer,

tout en ne ratant pas l'occasion de voir un bon film. Par exemple, ils évaluent si le fait que votre film soit court est un bon argument, ce qui guidera par exemple leur attention vers ce qui est prévu pour le reste de la soirée afin de décider si c'est une bonne raison. Mais vous aussi, quand vous essayez de convaincre, vous évaluez des raisons. D'une part pour contre argumenter : si vos amis vous répondent qu'ils se moquent que votre film soit court ou long puisque rien n'est prévu après, c'est l'évaluation de cette dernière raison qui vous guide. Soit, si vous la trouvez mauvaise, votre attention est orientée vers la recherche de représentations montrant qu'il y a des choses prévues après. Soit, si vous la trouvez bonne, votre attention est guidée vers la recherche d'autres arguments pour votre film. Dans ce dernier cas vous venez non seulement d'accepter l'argument, mais également l'état du monde qu'il contient : il n'y a rien de prévu après. Vous avez d'ailleurs l'intuition que l'acceptation de l'état « il n'y a rien de prévu après » fait certes échouer votre argument du film court, mais cela reste cohérent avec le choix de votre film. Vous cherchez et testez alors d'autres raisons qui pourraient faire pencher la balance de votre côté. Bien entendu, il est possible qu'au cours de la discussion l'évaluation de raisons pour choisir d'autres films vous convainque et vous fasse changer d'avis. Comme toute capacité communicative, le raisonnement a deux versants : nous sommes tous tantôt émetteur tantôt récepteur.

Si l'évaluation et la production de raisons sont liées du point de vue de leur fonction et de leurs contextes d'activations elles sont, d'un point de vue algorithmique, très liées également. Prenons une analogie avec le système d'attention visuelle qui peut être utilisé de façon « top-down », descendante ou « bottom-up», ascendante.

Utilisé de façon ascendante, le module traitant les visages par exemple détecte efficacement (même trop) la présence de pattern ressemblant à un visage. Utilisé de façon descendante, ces mêmes mécanismes permettent de rechercher un visage particulier dans une foule de visages. Dans ce dernier cas, votre système attentionnel garde les propriétés du visage pour les utiliser comme détecteur dans votre recherche. Dès que ces propriétés sont activées, c'est le bon visage.

Mais les propriétés d'un visage connu sont bien mieux définies que celles d'une raison qui « pourrait marcher » dans une conversation. Pour prendre une meilleure analogie que les visages dans le cadre de l'attention visuelle, prenons plutôt la recherche d'un outil improvisé (voir Barsalou 1983 sur les catégories ad hoc et Mercier 2009 pour l'analogie avec la recherche d'arguments).

Dans le cas descendant, imaginez-vous dans une salle pleine d'objets en tout genre et avoir pour objectif de trouver le plus rapidement possible un « outil pour tuer une guêpe posée sur une lampe en cristal fin ». L'objet doit être ni trop dur ni trop mou, par habitude vous pensez à un journal mais vous n'en voyez pas, vous scannez les objets à la recherche de ce qui pourrait convenir. Vous

pouvez dans les deux cas faire des essais, quitte à arrêter votre geste juste avant de casser la lampe avec un dictionnaire.

Pour convaincre, vous avez également la possibilité de faire des tests. Vous « testez » les premières raisons qui pourraient fonctionner auprès de votre auditoire. Au-delà de l'exemple où vous risquez de casser une lampe, les tests réels dans la nature sont généralement bien plus risqués à réaliser, alors que dans le cadre d'une discussion argumentative, il n'y a en général pas de conséquence trop fâcheuse à donner un argument qui s'avère non-convainquant. Comme nous l'avons déjà vu, vous pouvez ensuite soit chercher un argument qui rend votre argument convaincant, soit reprendre la recherche précédente pour un nouvel argument ; avec un peu de chance, vous avez même appris un peu plus d'informations au fil de la discussion sur ce qui pourrait faire mouche avec votre auditoire dans cette situation.

De façon ascendante, vous détectez les liens entre les représentations et inférez leur force immédiatement. Dans le cas de notre exemple de recherche d'outil improvisé, l'équivalent serait d'avoir développé la capacité d'évaluer immédiatement si un objet peut faire l'affaire dans les situations où un objet fragile est impliqué par exemple. Mais une différence dans le cas du raisonnement c'est la possibilité d'emboitement. Dans les cas où vous ne pouvez inférer intuitivement la force d'un lien raison-conclusion, vous inférez au moins une direction pour pouvoir l'évaluer. En effet, nous avons vu que vos mécanismes sont taillés pour scruter la cohérence de votre interlocuteur. Selon la pertinence de la conclusion et ce que l'émetteur veut faire de cet argument dans la conversation il peut être intéressant, pour vous, de creuser un peu et demander à l'émetteur de déployer plus de liens raison-conclusion pour pouvoir évaluer le lien précédent.

4.4 Que fait-on avec le raisonnement ?

Nous avons jusque-là présenté le raisonnement dans des exemples qui correspondent parfaitement à sa fonction : échanger des raisons en dialogue. Mais de la même façon que le module de détection de visage traite aussi les portraits, les masques, ou même s'active à chaque fois que deux points et un arc de cercle sont saillants, le module du raisonnement a un domaine propre et un domaine actuel (Sperber et Hirschfeld dans Carruther, Lawrence, Stich (Eds.), 2007). Dans le cas du raisonnement, les domaines actuels les plus proches du domaine propre sont l'explication et la prise de décision individuelle.

L'utilisation de raisons pour expliquer ou évaluer des explications est souvent considérée comme au moins aussi importante que l'usage de raisons pour justifier ou évaluer des justifications (voir par exemple Lombrozo 2011). Dans le cadre de la théorie argumentative en revanche, le module traitant de raisons a en partie évolué pour traiter d'explications d'actions (des justifications) mais peut recruter en dehors de son domaine propre, pour évaluer des explications en tout genre. Dans les deux cas, ce sont des liens raison-conclusion entre des représentations qui sont traités. Les travaux sur l'importance de l'explication s'appuient souvent sur des données montrant que les jeunes enfants sont capables de discriminer les bonnes explications des mauvaises. Pour prendre un exemple typique de cette jeune littérature tiré de Corriveau et Kurkul (2014).

Pour une question comme « Pourquoi pleut-il ? » Les enfants de cinq ans préfèrent une explication non-circulaire du type : « Il pleut car il y a des nuages dans le ciel qui sont remplis d'eau. Quand il y a trop d'eau dans les nuages, elle tombe sur le sol et nous rend tout mouillé » à des explications circulaires du type : « Il pleut car le temps est humide et nuageux et il tombe de l'eau du ciel. Lorsque l'eau tombe du ciel, on appelle ça la pluie et cela nous rend tout mouillé.» Or ces explications sont également des raisons pour accepter le fait qu'il pleuve. Pourtant la plupart des explications ne sont pas des raisons. Par exemple, à partir d'une affirmation comme « Jean s'est cassé la jambe », l'énoncé « Je l'ai vu avec un plâtre hier » est une raison pour accepter la première phrase mais pas une explication, alors que « Il est tombé de vélo » est une explication et une raison. Les jeunes enfants sont-ils aussi capables de distinguer une bonne explication d'une mauvaise lorsque celles-ci ne sont pas des raisons ? Sans investigations expérimentales supplémentaires séparant les deux, la question reste ouverte.

L'autre domaine actuel du module de raisonnement est la réflexion personnelle. Par exemple lorsque vous cherchez des arguments pour vos intuitions, anticipant un dialogue avec d'autres personnes qui pourraient ne pas les partager. Ou encore, avant de prendre une décision sur la base d'une intuition, quand vous pensez aux justifications possibles de votre action.

C'est peut-être ici la prédiction la plus intéressante de la théorie argumentative en comparaison avec les visions intellectualistes du raisonnement. Lorsque nous avons des intuitions relativement faibles concernant une décision, ce serait la recherche de justifications qui guiderait nos décisions et non simplement la recherche de la solution directement la plus optimale pour nous. Cette idée que les justifications d'une action guident nos décisions est celle proposée par Shafir et Tversky pour expliquer pourquoi, dans leur expérience, certains étudiants ne prennent pas la décision de partir en vacances sans connaître leurs résultats d'examen, alors qu'ils seraient partis dans tous les cas. Dans le cadre théorique du « reason-based choice » (voir Shaffir 1993 pour une

revue), lorsque les étudiants connaissent les résultats de leurs examens, ils ont facilement en tête des raisons justifiant la décision de partir trois jours en vacances (soit comme récompense, soit pour se changer les idées après une déception). Or, même si ces raisons pointent vers la même décision, elles sont quasiment incompatibles. Les participants n'ayant pas reçu leurs résultats se retrouvent donc à hésiter car leurs raisons pour partir en vacances sont perçues comme antagonistes.

Une des utilisations du module du raisonnement est donc l'aide à la prise de décision individuelle lorsque nous n'avons pas d'intuitions fortes pour prendre une décision.

Cela dit, même dans des cas où nous avons une intuition forte, le raisonnement dans la vision argumentative pourrait également, comme prédit par les théories intellectualistes, corriger nos intuitions initiales : s'il apparaît qu'il n'y a vraiment aucune justification acceptable pour une action envisagée. Aussi bénéfique que puisse être une décision pour vous, si les dégâts réputationnels qu'elle engendre sont trop importants, mieux vaut choisir une option qui vous apporte moins directement mais coûte moins sur le long terme. En ce sens la recherche de raisons peut améliorer la prise de décision individuelle car elle permet de vérifier que les coûts réputationnels d'une action ne seront pas trop grands.

4.5 Un type de processus, deux types d'inférences

Concluons ce chapitre sur la théorie argumentative du raisonnement en mettant en évidence une première tension entre les théories intellectualistes du raisonnement et cette théorie interactionniste. Dans la vision classique le raisonnement est ce qui nous permet de passer des prémisses aux conclusions. De passer de « tous les hommes sont mortels » à « tous les Grecs sont mortels ». Dans la vision interactionniste le raisonnement semble toujours aller à rebours, cherchant des arguments supportant une conclusion déjà acquise.

Cependant, lorsque nous évaluons une raison comme suffisamment bonne, nous acceptons du même coup l'information pour laquelle le récepteur vient de produire cette raison. Dit autrement, ce que produit directement notre module de raisonnement est une intuition sur le lien entre une raison et une conclusion. Dans les cas où ce lien serait jugé comme suffisamment fort, le produit indirect du module de raisonnement est l'acceptation de la conclusion elle-même. Lorsque nous acceptons une conclusion à partir d'une intuition sur une raison, nous produisons ce que nous appellerons une inférence réflexive en opposition aux inférences intuitives pour lesquelles les raisons qui nous ont menés à la conclusion échappent à notre conscience. Prenons des exemples afin d'illustrer le contraste entre inférences intuitives et inférences réflexives.

Remarquons tout d'abord que la plupart de nos inférences sont intuitives. Lorsque vous inférez l'émotion que ressent quelqu'un à partir de certains traits sur son visage, lorsque vous inférez qu'il va pleuvoir en voyant un gros nuage noir, ou encore lorsque vous inférez que quelqu'un approche en entendant des bruits de pas, vous n'avez aucun accès conscient à ce qui vous a mené de la prémisse aux conclusions de ces inférences.

Dans le cas des inférences réflexives en revanche, la raison qui vous a mené de la prémisse à la conclusion est représentée, au moins minimalement. Vous allez regarder un match de votre équipe de football favorite, vous remarquez que la star de l'équipe ne joue pas et inférez que votre équipe va perdre. Après un calcul, vous obtenez une probabilité de 1.2 et inférez que votre calcul doit être faux. Outre le fait qu'elle repose sur des raisons, les inférences réflexives sont similaires aux inférences intuitives, elles sont généralement rapides, peu coûteuses et comportent une dimension inconsciente au sens où nous n'avons pas automatiquement accès au contenu de ces raisons. Nous savons cependant qu'il y en a.

Des inférences réflexives peuvent également être emboitées les unes dans les autres. Pour illustrer ce point, imaginez être un sujet ayant répondu 0.10€ au problème du bonbon et de la baguette. Un inconnu arrive alors et vous dit que :

(1) « la réponse n'est pas 0.10€ »

Au vu de votre regard dubitatif il vous explique :

(2) « Si le bonbon valait 0.10€, le tout vaudrait 1.20€ ».

Vous froncez les sourcils et la personne explique alors les raisons de son inférence :

(3) « La baguette vaut 1€ de plus que le bonbon, donc si le bonbon vaut 0.10€, la baguette vaut 1.10€. Le tout ferait donc 1.20€ »

On peut imaginer qu'à ce moment-là, (3) soit suffisamment intuitif pour que vous acceptiez la phrase (2), mais cela pourrait ne pas être le cas, l'inconnu rentrerait alors dans les raisons pour (3), par exemple en faisant le calcul avec vous (voire même avec l'énoncé sous les yeux). Admettons donc que vous ayez accepté (2) en comprenant (3), vous comprenez maintenant les raisons pour accepter (1) et admettez que « la réponse n'est pas 0.10€ ».

Même si vous n'avez pas en tête toutes les raisons emboitées qui vous ont fait accepter (1), vous savez qu'elles existent. Peu importe au bout de combien de niveau d'emboitement cela vous a paru intuitif, vous avez intuitivement évalué comme bonnes les raisons pour (1), et vous décidez donc d'accepter cette conclusion.

Le fait que des raisons existent entre une prémisse et une conclusion n'est cependant pas synonyme d'inférence réflexive. Lorsque vous percevez la phrase mathématique « 1+1 », « 2 » vous vient immédiatement en tête. Vous n'avez aucun accès conscient à ce qui vous a mené de l'un à l'autre. Pourtant, il existe une série d'arguments mathématiques complexes permettant de justifier cette inférence à partir de quelques axiomes arithmétiques plus ou moins intuitifs. Il y a moins d'arguments emboités pour justifier le fait que « la star de l'équipe ne joue pas » vous ait fait inférer « l'équipe va perdre », qu'il n'y en a pour justifier que « 1+1 » vous ait fait inférer « 2 ». Pourtant la première inférence est réflexive et la deuxième est intuitive.

Nous avions dans le premier chapitre relevé une apparente contradiction entre le fait que le raisonnement soit une capacité générale et qu'il puisse être un module spécialisé. La théorie argumentative apporte une solution à ce paradoxe.
Le raisonnement est un module méta-représentationnel spécialisé dans les raisons. Un module méta-représentationnel est un module qui porte sur des représentations, en opposition à la plupart des autres modules (comme celui traitant des visages) qui portent sur des objets du monde extérieur. Un module représentant des représentations peut traiter de représentations qui portent sur n'importe quels objets du monde, ce qui explique sa généralité. Prenons une analogie avec les représentations des nombres, pouvant référer à une quantité de n'importe quoi dans le monde.

Lorsque vous avez l'intuition que 10 fois 100 égal 1000, vous utilisez une propriété sur les représentations des quantités, ici le fait que vous soyez en base 10. Pour quelqu'un qui aurait une représentation différente des quantités, les chiffres romains par exemple, le fait que X fois C fait M n'est pas très intuitif. Que cela se réfère à 10 personnes par jour pendant 100 jours, ou à 10 cordes de 100 mètres, pour arriver à 1000, vous tirez parti des représentations que vous avez de ces objets du monde, pas des objets eux-mêmes. De la même façon, le module du raisonnement nous donne des intuitions sur les liens raison-conclusion entre nos représentations, ces représentations pouvant porter sur n'importe quoi. Le module de raisonnement est donc à la fois très spécialisé – s'occupant uniquement de liens raison-conclusion – et en même temps d'une grande généralité. En opposition à des mécanismes « généraux » au sens entendu classiquement, Mercier et Sperber (2009) parlent, pour les modules méta-représentationnel, de généralité virtuelle.

En conclusion de ce chapitre, mettons en avant ce que la théorie argumentative a de plus différent par rapport aux théories à double processus d'un point de vue algorithmique. Le raisonnement est essentiellement un type de processus « système 1 ». C'est un module inférentiel qui nous donne des intuitions sur les liens raison-conclusion de nos représentations.

Nous allons voir, dans le chapitre suivant, comment cette caractérisation du raisonnement comme mécanisme argumentatif peut expliquer les résultats de la psychologie du raisonnement. Nous comparerons son pouvoir explicatif avec celui des théories intellectualistes dont, en particulier, les théories à processus duels.

Chapitre 5 – Les échecs du raisonnement individuel

Comment la théorie argumentative rend-elle compte des performances observées par les psychologues du raisonnement ?

Tout d'abord, pour beaucoup de tâches étudiées dans le domaine du raisonnement, comme des problèmes de logique abstraits par exemple, on peut reprendre une analogie de (Mercier 2009).

« 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. Chercher des explications sur pourquoi si peu de sujets arrive à marcher sur les mains ne nous apprendra pas grand chose sur les capacités de marche humaine ».

Si on ne fait pas d'hypothèse logiciste l'échec des sujets à respecter les lois logiques est aussi surprenant que l'échec des sujets à marcher sur les mains. Cependant, cela ne veut pas dire que tous les échecs des sujets dans les tâches classiques de la psychologie du raisonnement s'expliquent simplement par le fait que les lois de la logiques, ou plus largement normatives, soient en elles-mêmes non-pertinentes pour notre esprit.

Revenons à présent sur le problème de l'activation du raisonnement dans le cadre des théories à processus duels. Les cas les plus problématiques pour ces théories sont, de notre point de vue, les problèmes de raisonnement incongruents pour lesquels la bonne réponse est accessible aux sujets. Par exemple le problème du bonbon et de la baguette ou encore celui de Paul et Linda. En effet, si la fonction du raisonnement est de corriger nos intuitions lorsqu'elles sont trompeuses, ces problèmes semblent être des cas idéaux pour l'activation du raisonnement. Nous avons ébauché les explications que fournissent les théories à processus duels pour ces échecs du raisonnement, la principale étant le manque de ressources cognitives. Nous allons voir à présent pourquoi un mécanisme intuitif de production et d'évaluation de raison peut expliquer ces échecs du raisonnement individuel.

5.1 Le double échec du raisonnement individuel

Avant d'expliquer ces échecs dans le cadre de la théorie argumentative du raisonnement, remarquons que les erreurs des sujets sont de deux natures différentes : non seulement la majorité des sujets échoue à donner la bonne réponse mais, de plus, les sujets qui donnent la mauvaise réponse expriment une confiance remarquablement forte dans leur mauvaise réponse. Afin d'illustrer cette idée présentons les premières données expérimentales originales de cette thèse.

Pour leur premier cours de première année de Licence en sciences cognitives à l'Université de Lyon, 226 étudiants ont été les participants d'une expérience. Tous ont eu à résoudre deux problèmes : celui du bonbon et la baguette et celui de Paul et Linda.

Le vrai but cet expérience était social, mais nous ne présenterons ici que les données de la première phase en individuel. Observons l'évolution du taux de bonne réponse et de la confiance des sujets en leur réponse au cours du temps. Une fois le premier problème affiché au tableau, les sujets devaient essayer de résoudre le problème seul et fournir une réponse avec leur sentiment de confiance de 1 à 10, 20 secondes après

(Temps 1) 1minute et 20s après (Temps 2), 2 minutes et 20s après (Temps 3), 3 minutes et 20s après (Temps 4) et enfin 4 minutes et 20s après (Temps 5). Ensuite les sujets passaient à la phase sociale de l'expérience que nous décrirons dans le chapitre suivant. Puis, les participants recommençaient la même chose avec l'autre problème.

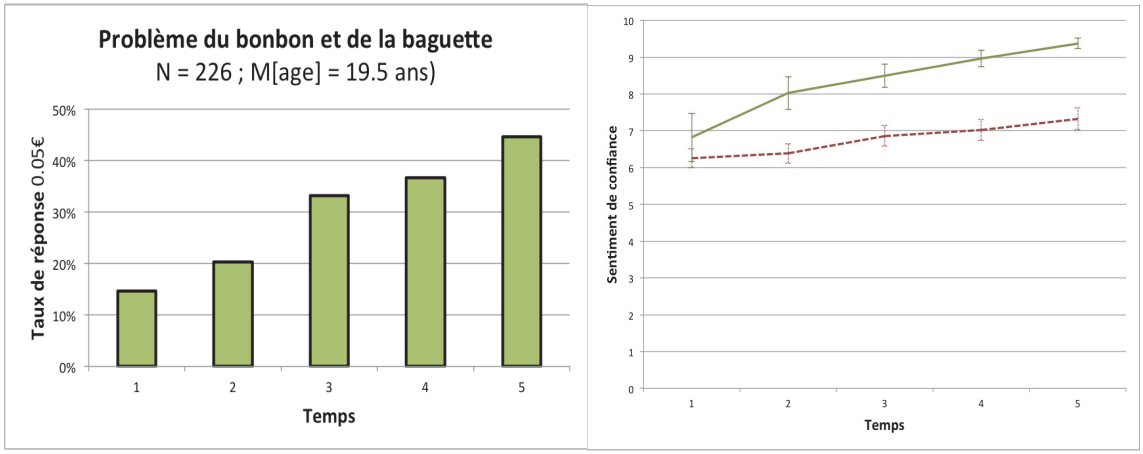


Figure 1 A gauche l'évolution du taux de réponse 0.05€. A droite l'évolution du sentiment de confiance moyen pour les sujets qui répondent 0.05€ (vert) et 0.10€ (pointillé). Les bars indiquent les erreurs-standard.

On observe qu'en l'espace de quatre minutes et vingt secondes, de plus en plus de sujets ont corrigé leur mauvaise intuition initiale. On pourrait penser que, si on laissait plus de temps aux sujets, tout le monde aurait la bonne réponse mais, comme nous le verrons par la suite, c'est loin d'être le cas. Ce qui nous intéresse ici est que le sentiment de confiance augmente au cours du temps, que les sujets aient la réponse correcte ou incorrecte. Comme l'indiquent les tests de rangs signés de Wilcoxon sur les sujets qui ne changent pas d'avis entre les temps 1 et 5, les 32 sujets qui gardent la réponse correcte passent d'une confiance moyenne de 7.0 (SD = 3.7, Mdn = 9) à une confiance de 9.7 (SD= 1.0, Mdn = 10), $Z = -3.5$, $p < .001$, $r = 0.44$. Et pour les 124 sujets gardant la mauvaise réponse, ils passent d'une confiance moyenne de 6.4 (SD = 3.6, Mdn = 7) à une confiance de 7.4 (SD = 3.3, Mdn = 9), $Z = -2.6$, $p = .009$, $r = 0.16$.

Si le raisonnement semble bien corriger l'intuition initiale de certains sujets, il semble aussi rendre tout le monde plus confiant. Enfin, remarquons que les sujets qui ont la mauvaise réponse sont moins confiants que ceux qui ont la bonne réponse, ce qui peut suggérer, de façon cohérente avec les données de De Neys et collaborateurs, que ces

sujets dans l'erreur ont un signal inconscient que quelque chose n'est pas normal. La comparaison expérimentale est cependant moins rigoureuse ici car la comparaison se fait avec des sujets qui sont passés par l'intuition mais ont changés d'avis. Par ailleurs la confiance des sujets dans l'erreur, plus basse au départ, augmente moins vite que pour les sujets qui ont la bonne réponse dont la majorité est au plafond. Nous essaierons dans la suite d'expliquer ce phénomène de sur-confiance.

Ces sujets, qui pour l'immense majorité revenaient tout juste de vacances après leur réussite au baccalauréat, ont un taux de réponse relativement haut par rapport aux observations de la littérature. Avant de présenter leurs résultats au problème de Paul et Linda, présentons des données utilisant exactement le même paradigme, mais pour une population de professionnel de la santé, majoritairement infirmiers et éducateurs (M[age] = 42.9, SD = 6.8). La seule différence avec le design expérimental précédent est que la première réponse est demandée aux sujets une minute après la présentation du problème et non vingt secondes après. Comme précédemment, il leur est ensuite demandé de répondre toutes les minutes, jusqu'à la quatrième minute. Voici les résultats.

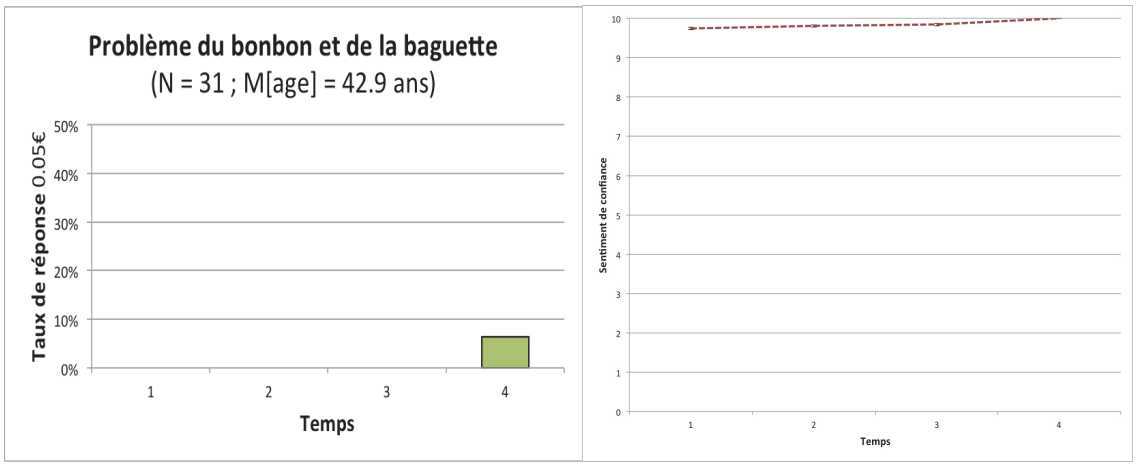


Figure 2 A gauche l'évolution du taux de réponse 0.05€. A droite l'évolution du sentiment de confiance moyen pour les sujets qui répondent et 0.10€ (pointillé), les deux sujets qui trouvent la bonne réponse au temps 4 n'apparaissent pas.

Seul deux sujets finissent par trouver la bonne réponse au bout de quatre minutes de raisonnement individuel. Le sentiment de confiance des sujets est cette fois quasiment au plafond dès la première minute. Pour les deux sujets qui finissent avec la bonne réponse leur confiance passe du maximum pour la mauvaise réponse, au maximum pour la bonne

réponse. Pour le reste de cette population le raisonnement semble, non seulement ne pas corriger la mauvaise intuition, mais ne pas laisser place au doute non plus, les sujets étant quasiment au plafond dès la première minute.

Revenons maintenant à nos 226 étudiants en première année de licence, et présentons leurs résultats à la tâche disjonctive de Paul & Linda.

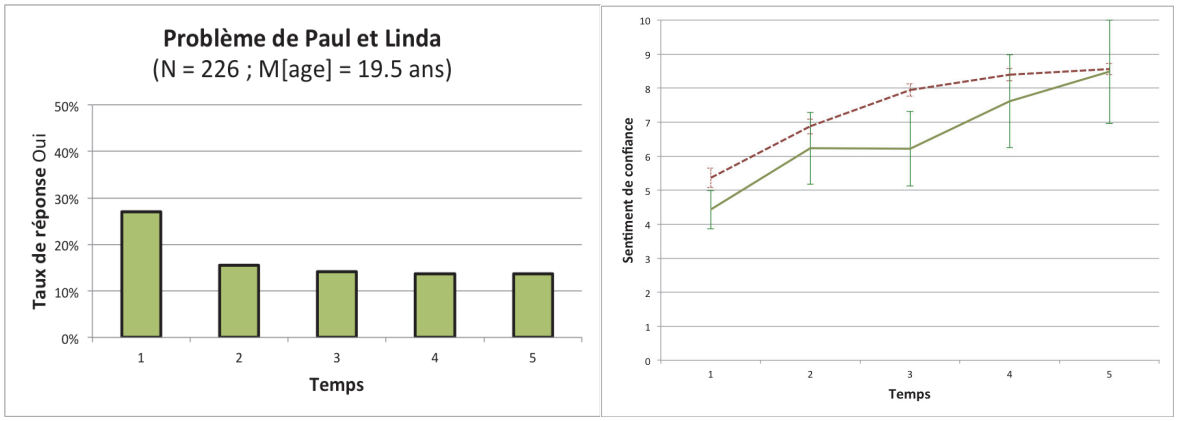


Figure 4 A gauche l'évolution du taux de réponse « Oui ». A droite l'évolution du sentiment de confiance moyen pour les sujets qui répondent « Oui » (en vert) et « On ne peut pas savoir » (en pointillé).

Si dans le cas du bonbon et de la baguette, le taux de réponse "0.05€" indique parfaitement le taux de bonne réponse, le problème de Paul et Linda est un problème qui n'a que trois réponses possibles. Oui, Non, et On peut pas savoir. Ainsi, certains sujets peuvent répondre la bonne réponse "oui" mais pour de mauvaises raisons, ce qui peut difficilement être considéré comme une bonne réponse. Ceci explique l'allure du premier graphique entre les temps 1 et 2, qui ne montrent pas un taux de réussite en baisse, mais simplement des sujets changeant de mauvaise réponse. La suite de l'expérience que nous ne présenterons que plus tard nous permet cependant de penser que les 14% de sujets qui finissent par répondre Oui au bout de quatre minutes et vingt secondes de raisonnement le font avec la bonne justification, et ont donc la bonne réponse.

Comme dans le cas du problème du bonbon et de la baguette, le sentiment de confiance augmente minute après minute pour tous les sujets. Les sujets qui ont la mauvaise réponse ne semblent pas vraiment moins confiants que ceux qui ont la bonne réponse, mais il est difficile de tirer des conclusion avec ces données car la population de

sujets qui répond "oui" pourrait contenir des sujets qui n'ont, en fait, pas vraiment la bonne réponse au problème.

Pour savoir si les sujets qui donnent la mauvaise réponse au problème de Paul et Linda sont moins confiants que les sujets qui donnent la bonne réponse, présentons des données récoltées en ligne grâce à la plateforme Amazon Mechanical Turk.

Nous avions réalisé diverses expériences avant le début de cette thèse utilisant le problème de Paul et Linda. Présentons ici un agrégat de données de plusieurs expériences ayant la même phase individuelle. Les sujets doivent répondre au problème, justifier leur réponse (ce qui nous permet de savoir qui a vraiment la bonne réponse), puis donner leur sentiment de confiance. Comparons le sentiment de confiance pour deux populations de sujets : les sujets qui ont la bonne réponse avec la bonne justification, et les sujets qui donnent la réponse intuitive au problème ("On ne peut pas savoir"). En tout, il y a 442 sujets (Mage = 33.8, SD = 11.9) dont seulement 10% donnent la bonne réponse. La confiance moyenne mesurée de 0 à 9 et normalisée en pourcentage est de 74.4% (SD = 23.0) pour des sujets qui ont la bonne réponse et de 73.0% (SD = 23.0) pour ceux qui donnent la réponse intuitive. Une différence non significative, test de Wilcoxon Mann - Whitney ($Z = -0.4$, $p = 0.69$, $r = 0.02$). Par rapport au problème du bonbon et de la baguette, les sujets qui donnent la réponse intuitive au problème de Paul et Linda ne semblent pas détecter qu'ils se trompent. Nous avons des confirmations de cette non différence avec des données récoltées chez des étudiants, la différence moyenne va toujours dans le même sens mais toujours très loin d'être significative.

En conclusion, une théorie du raisonnement doit non seulement expliquer le fait que le raisonnement ne corrige pas les mauvaises intuitions initiales, mais également pourquoi le raisonnement semble mener les sujets à être sur-confiants, même dans d'erreur. C'est d'autant plus frappant pour des problèmes comme les deux présentés car la réponse pourrait être parfaitement accessible aux sujets.

Nous avons donné quelques éléments fournis par les théories intellectualistes du raisonnement pour expliquer que le raisonnement semble très souvent ne pas accomplir sa fonction supposée : corriger nos intuitions initiales. Il est difficile d'imaginer comment ces théories pourraient expliquer le phénomène de sur-confiance. Nous allons maintenant

tenter d'expliquer ce double échec du raisonnement à la lumière de la théorie argumentative du raisonnement.

5.2 Le biais vers son côté

La première explication vient d'un des phénomènes les plus robustement observés en psychologie, connu sous le nom de "biais de confirmation". Nous allons défendre l'idée que ce phénomène devrait changer de noms pour "biais vers son côté" mais, avant, présentons des expériences ayant tenté de mettre en évidence le phénomène.

Deanna Kuhn, pionnière dans l'étude de l'argumentation et de la cognition, a demandé à 160 participants de prendre position sur diverses questions comme l'échec scolaire ou le chômage. Une fois que les participants avaient donné leur avis il leur était demandé de se justifier. Presque tous les participants ont facilement produit des raisons pour soutenir leur point de vue. Mais quand il leur a été demandé de produire des contre-arguments à leur propre point de vue, seulement 14% étaient toujours en mesure de le faire, la plupart d'entre eux préférant ne rien produire du tout (Kuhn 1991).

On ne peut cependant pas considérer, comme certains auteurs (Edwards & Smith 1996), que ce type d'expérience est une preuve expérimentale concluante du biais de confirmation. En effet, le fait de pouvoir donner plus facilement des arguments pour son point de vue plutôt que contre pourrait tout simplement venir du fait que, si nous avons ce point de vue, c'est précisément parce que nous avons plus de raisons le soutenant.

D'autres expériences ont cependant démontré de façon quasiment parfaite l'existence du biais de confirmation en utilisant un paradigme de « cécité au choix » (*choice blindess*). Peter Johansson et ses collaborateurs ont interrogé des passants sur une série d'affirmations politiques, comme "La surveillance à grande échelle de nos e-mails et du trafic internet par le gouvernement devrait être interdite comme moyen de lutte contre le terrorisme international" ou des affirmations morale comme "si une action peut heurter une personne innocente, il est moralement répréhensible de la réaliser". Les sujets devaient, pour chaque phrase, donner leur avis entre 1 - Pas du tout d'accord et 9 - complètement

d'accord. Dans la deuxième phase de l'expérience l'expérimentateur proposa aux sujets de revenir sur trois de leurs jugements pour qu'ils les expliquent. Par un tour habile de passe-passe certaines des affirmations avaient cependant été remplacées par l'affirmation inverse. Pour reprendre les exemples précédents, les affirmations étaient devenues "La surveillance à grande échelle de nos e-mails et du trafic internet par le gouvernement devrait être autorisée comme moyen de lutte contre le terrorisme international" et "si une action peut heurter une personne innocente, ce n'est pas forcément moralement répréhensible de la réaliser". Les jugements initiaux de 1 à 9 des sujets, eux, étaient restés inchangés. 69% des participants ne détectèrent pas la manipulation pour au moins une des affirmations. Même parmi les participants ayant choisi un extrême de l'échelle, 31% ne se rendirent compte de rien. Ces participants "non-détecteurs" se retrouvèrent alors parfaitement capables de justifier une position pour laquelle ils avaient choisi la position inverse quelques minutes auparavant (Hall, Johansson, & Strandberg, 2012).
Cette démonstration ingénieuses du biais de confirmation illustre parfaitement notre capacité à justifier n'importe quel point de vue que l'on pense être le notre.

Expliquons à présent pourquoi cette capacité à justifier nos positions n'est pas un biais de confirmation mais bien un biais vers son côté. Tout simplement car, comme (Shaw 1993) le montre, lorsque il est demandé aux participants de produire des arguments contre un point de vue avec lequel ils sont en désaccord, ils n'ont aucun mal à produire des arguments infirmant ce point de vue. Nous parlerons de biais vers son côté (*myside bias*).

Le fait que notre raisonnement souffre d'un biais vers son côté est une prédiction directe de la théorie argumentative du raisonnement. Ce n'est d'ailleurs plus un « biais » dans un contexte argumentatif. Si vous devez convaincre quelqu'un que votre intuition est la bonne, il suffit de trouver des raisons pour votre point de vue, pas pour celui de votre interlocuteur. Votre interlocuteur se chargera bien lui-même de chercher des arguments pour son point de vue. Dans ce cadre le biais vers son côté est un mécanisme parfaitement adapté aux dialogues argumentatifs. Il représente une division du travail cognitif, chacun cherchant des arguments pour son point de vue. En échangeant ces raisons, les meilleures pourront l'emporter.

Pour les théories intellectualistes, en revanche, l'existence du biais vers son côté est un vrai problème. Si la fonction du raisonnement est de corriger nos intuitions trompeuses, le fait qu'il ait tendance à produire des représentations supportant nos intuitions initiales est une caractéristique qui semble aller directement à l'encontre de sa fonction. A notre connaissance aucun psychologue ne nie son existence, et les partisans des théories intellectualistes du raisonnement s'accordent à dire que c'est un trait qui nuit à la rationalité. Comme le résume Raymond Nickerson, dans sa revue de littérature sur le sujet :

« La majorité des commentateurs, de loin, voient le biais de confirmation comme un échec humain, une tendance qui est à la fois omniprésente et irrationnelle. Ce n'est pas difficile de défendre cette position. Le biais peut contribuer à des illusions de toutes sortes, au développement et à la survie des superstitions et à une variété d'états indésirables de l'esprit, y compris la paranoïa et la dépression. Il peut être exploité par les voyants, devins, diseurs de bonne aventure, ou par toute personne ayant un penchant pour faire adopter à d'autre des allégations non fondées. On peut aussi imaginer qu'il joue un rôle important dans la perpétuation des animosités et des conflits entre personnes ayant des vues contradictoires sur le monde »

(Nickerson 1998, p. 205)

Comment une telle caractéristique du raisonnement peut-elle être conciliable avec les théories intellectualistes et l'idée que la fonction du raisonnement est de corriger nos intuitions ? A notre connaissance aucune stratégie pour sortir de cette impasse n'est satisfaisante. Le biais vers son côté ne rend pas simplement le rôle supposé de correction du raisonnement plus difficile, il va directement à son encontre.

Par opposition aux mécanismes de test d'hypothèse décrits par Evans et Stanovich, plutôt que d'évaluer chaque option de façon à peu près équilibrée pour espérer prendre de meilleures décisions, le biais vers son côté nous fait produire des raisons pour notre point de vue, ou contre des points de vue différents. Reprenons l'exemple dont nous nous sommes servi pour illustrer la théorie d'Evans et Stanovich. Si à partir du moment où vous avez l'intuition d'amener votre fils à la chasse, votre raisonnement ne cherche que des raisons pour cette intuition, il ne vous reste plus qu'à espérer pour votre enfant (donc pour vos gènes) que ce soit une bonne première intuition.

Les partisans des théories intellectualistes sont bien conscients du problème que pose le biais vers son côté pour leur théorie. Stanovich par exemple, dans son livre « *The Robot's Rebellion* » (2004) classe ce biais dans la liste des biais du système intuitif. Expliquons brièvement pourquoi cela ne tient pas.

Premièrement le biais vers son côté n'est présent chez aucun animal non-humain et pour cause : des mécanismes intuitifs préfèrent la surprise à la confirmation. La surprise, au sens d'erreur de prédiction, est l'ingrédient essentiel pour mettre à jour les attentes de ces mécanismes. Si nos mécanismes intuitifs sont biaisés, c'est vers la surprise et non vers la confirmation.

Deuxièmement Thomas Allen et collaborateurs ont offert une démonstration expérimentale des biais respectifs de nos intuitions et de notre raisonnement. Une image d'un individu était montrée aux participants avec deux énoncés décrivant le comportement de cet individu. Il était ensuite demandé aux sujets de donner leurs impressions concernant l'individu. Certains participants ont vu l'image d'un «jeune adulte mâle noir portant un bandeau noir et lunettes noires" suivi de deux déclarations: "Impoli avec la vendeuse» et «A laissé sa place à une personne âgé dans un métro bondé". Si vous êtes familier avec les stéréotypes des jeunes hommes noirs aux États-Unis, vous avez compris que la première déclaration a été conçue pour aller dans le sens des attentes de la plupart des sujets, tandis que la seconde a été conçue pour être surprenante. Afin de tester le rôle de l'intuition dans la formation d'impression, la moitié des participants a été empêchée de raisonner, ayant à retenir une longue série de chiffres. Ces participants ont été guidés par leurs intuitions et ont plus porté leur attention sur la déclaration surprenante. En revanche, les participants qui pouvaient raisonner ont plus porté leur attention sur la déclaration « non-surprenante ». Les intuitions viseraient donc à recueillir les informations les plus utiles, alors que le raisonnement viserait à confirmer les stéréotypes des participants.

5.3 La paresse sélective du raisonnement

Le biais vers son côté n'est cependant pas le seul problème du raisonnement. Le raisonnement semble également paresseux lorsqu'il produit des arguments, se contentant souvent d'arguments de faible qualité. Ce fait fut observé par un des pionniers des études sur l'argumentation, David Perkins. Dans un design similaire à l'étude de Kuhn présentée précédemment, (Perkins 1985) demande aux sujets de justifier leur position sur des questions de société comme « *Would restoring the military draft significantly increase America's ability to influence world events?* ».

Perkins fait alors l'observation que les arguments des sujets sont superficiels « […] les raisonneurs analysent une situation seulement de manière à ce que cette analyse ait du sens ». Il caractérise les sujets de « make sense epistemologist » (p. 568) indiquant qu'ils se satisfont d'arguments faibles pour leur position, ne faisant sens que superficiellement. De façon analogue, lors de la résolution de problèmes comme le bonbon et la baguette ou Paul et Linda, au vu de l'évolution de leur sentiment de confiance, les sujets semblent également parfaitement satisfaits de leurs arguments (nécessairement faibles) pour la mauvaise réponse. La faible qualité des arguments produits peut sembler problématique pour une théorie argumentative du raisonnement. En effet, si la fonction du raisonnement est de produire et évaluer des arguments, ne devrions-nous pas produire des arguments de qualité et les évaluer de façon critique ? Cela pourrait-il s'expliquer par nos limites en mémoire de travail, ou par la « paresse cognitive » de notre raisonnement.

Comme pour le biais vers son côté, les choses s'éclairent une fois de plus si l'on resitue le raisonnement dans le contexte pour lequel il a évolué : l'interaction dialogique. Lorsque vous essayez de convaincre vos amis de regarder un film, est-il nécessaire de démarrer une grande tirade expliquant tous les raisons pour lesquelles votre film est le meilleur choix ? De la même façon, est-il nécessaire d'essayer d'anticiper les goûts et les humeurs de chacun pour produire un argument de la plus grande qualité ? Une autre stratégie s'avère beaucoup moins coûteuse : commencer par le premier argument qui vous vient en tête et se servir du feedback de votre audience. Après tout, l'audience pourrait accepter ce premier argument, vous laissant ainsi parvenir à vos fins à moindre coût. Si le

premier argument ne fait pas mouche, vous n'avez qu'un faible coût à payer. Vous pourrez toujours essayer d'autres arguments. De plus, en évaluant les potentiels contre-arguments de votre interlocuteur, vous pourrez ensuite produire des arguments y répondant. Cette idée est développée dans le chapitre qui suit, rédigé en collaboration avec Hugo Mercier et Pierre Bonnier, publié dans « Cognitive Unconscious and Human Rationality » et intitulé « Why don't people produce better arguments?».

Nous présenterons ensuite une étude expérimentale montrant que le raisonnement n'est pas paresseux de façon générale mais fait preuve de ce que nous avons appelé une « paresse sélective ». Si nous évaluons les arguments qui sont en accord avec notre point de vue de manière superficielle, nous sommes en revanche critiques lorsqu'il s'agit d'arguments allant à son encontre. Avec Hugo Mercier et en collaboration avec les chercheurs suédois Petter Johansson et Lars Hall, j'ai utilisé un paradigme de cécité au choix pour faire évaluer aux sujets leurs propres arguments comme s'ils venaient de quelqu'un d'autre. Cela constitue une démonstration expérimentale de la « paresse sélective » : nous somme plus laxistes pour évaluer nos arguments que ceux des autres. Cette étude a fait l'objet d'une publication dans la revue Cognitive Science en 2015

Why don't people produce better arguments?

Mercier, Hugo[1]; Bonnier, Pierre[2] ; Trouche, Emmanuel

(1) Cognitive Science Center - University of Neuchâtel

(2) Ecole des Hautes Etudes en Sciences Sociales

Why don't people produce better arguments ?

That people produce arguments of low quality has been a recurring complaint from scholars of informal reasoning (Kuhn, 1991; Perkins, Farady, & Bushey, 1991), formal reasoning (Evans, 2002), and social psychologists (Nisbett & Ross, 1980). One of the main issues is that people tend to produce arguments that are one-sided (the one side always being their side) (Baron, 1995; Nickerson, 1998), and that they have trouble finding arguments for any other position (e.g. Kuhn, 1991). However, this is not the only problem: these biased arguments are often weak, making only "superficial sense," (Perkins, 1985, p. 568) as if people were content with the first argument that crosses their mind (Nisbett & Ross, 1980, p. 119).

These conclusions, reached in the study of arguments actually produced by participants, are bolstered by reasoning's failure to correct participants' intuitions in many tasks (e.g. Frederick, 2005; Wason, 1966). When people persist, after several minutes of reasoning, in providing the wrong answer to a simple logical or mathematical problem, it means not only that they mostly looked for arguments supporting their initial, wrong intuition, but also that they were satisfied with the arguments they found—arguments that were necessarily flawed given the nature of the tasks.

Moreover, recent research bearing on confidence in reasoning has revealed that participants are often very confident in their arguments, even when they are faulty (De Neys, Cromheeke, & Osman, 2011; Shynkaruk & Thompson, 2006). In particular, a study by Trouche et al (submitted) asked participants, for a standard reasoning problem, not only to evaluate the confidence in their answers, but also the confidence in the reasons for their answers. The participants who gave the intuitive but wrong answer were not only highly confident in their answer, but also in the—necessarily faulty—reasons for the answer.

The objective of this article is to explain why people seem to produce such weak arguments. In the first section, we lay out how two theories of the function of reasoning—the classical theory and the argumentative theory—account for reasoning's apparent limitations. The argumentative theory, we contend, can easily explain some of these apparent limitations, such as the myside bias, as well as more obviously adaptive features of reasoning, such as the ability to properly evaluate others' arguments. However, it is less clear how the argumentative theory can be reconciled with the low quality of arguments produced by reasoning.

Here we offer an explanation that rests on the dialogic nature of argumentation. When people produce arguments in a dialogue, they can rely on their interlocutor to explain why they find the argument defective or weak. Relying on interlocutor feedback is often more effective than trying to anticipate what a better argument would be. We then describe in more details how this account explains the various types of argument failures.

Two theories of reasoning and two features of argument production

It is useful to distinguish two well-established traits of argument production: the tendency to find arguments that support the reasoner's side, and the tendency to be satisfied by relatively weak arguments. The first of these traits has been the focus of intense study, generally under the name of confirmation bias (Nickerson, 1998). Although this research has soundly established that reasoning is biased, the word 'confirmation' is a misnomer: reasoning doesn't seek to confirm everything, only beliefs the reasoner shares. By contrast, when reasoning bears on beliefs the reasoner disagrees with, it produces counter-examples, counter-arguments and other ways to falsify the beliefs (see Mercier & Sperber, 2011). As a result, it is more accurate to talk of a *myside bias* (Mercier & Sperber, in prep).

The myside bias flies in the face of the classical theory of (the function of) reasoning. Most scholars who have speculated about the function of reasoning postulate that it has a chiefly individual function: to correct the reasoner's mistaken intuitions, thereby guiding her towards better beliefs and decisions (Evans, 2008; Kahneman, 2003; Stanovich, 2004). To perform this function properly, reasoning should either impartially look for reasons why the reasoner might be right or wrong or, even better, preferentially look for reasons she might be wrong. Reasoning does the exact opposite, behaving in a way that is difficult to reconcile with the classical theory of reasoning.

The second apparent limitation of reasoning—that it tends to produce relatively weak arguments—has been the focus of less intense scrutiny. Yet it is no less problematic than the myside bias. If people applied very high quality criteria to their own arguments, the effects of the myside bias would be much softened. In some cases—for instance in logical or mathematical tasks—people would have to admit that there are no good reasons for the intuitive but wrong answer, and they would be forced to change their mind. Again, the classical theory of reasoning

should predict the exact opposite: in order to make sure our intuitions do not lead us astray, reasoning should check that we have good reasons for them, not just any reason.

To reconcile these two traits of reasoning—the production of biased *and* weak arguments—with the classical theory of reasoning, psychologists often invoke cognitive limitations such as low working memory (e.g. Evans, 2008). However, cognitive limitations cannot be the main explanation for these traits, since reasoning only exhibits them when it produces arguments. When reasoning evaluates other people's arguments, it becomes (relatively) objective and exigent. It is objective because it accepts strong arguments and then leads the reasoner to change her mind. It is exigent because it rejects weak arguments. The good performance in reasoning tasks following group discussion demonstrates both traits of argument evaluation: if it wasn't (relatively) objective, people would reject the arguments for the good answer; if it wasn't (relatively) exigent, people would just as likely be convinced by arguments for the wrong answer (e.g. Laughlin & Ellis, 1986; Moshman & Geil, 1998). A large literature in social psychology also shows that when they care about the conclusion of an argument, people change their mind more in response to strong than to weak arguments (see Petty & Wegener, 1998). Other experiments have shown that participants are not easily swayed by straightforward fallacies, but that they react appropriately to sounder versions of the same arguments (Hahn & Oaksford, 2007).

Given that people are often careless in the evaluation of their own arguments—they produce relatively weak arguments—but that they judge other's people arguments more stringently, we suggest calling this property of reasoning *asymmetric argument evaluation*.

The argumentative theory of reasoning was developed as an alternative to the classical theory of reasoning (Mercier & Sperber, 2011). Instead of postulating an individual function for reasoning, it grounds the evolution of reasoning in the logic of the evolution of communication. Humans' reliance on communication creates selection pressures for mechanisms that protect receivers from the potentially harmful information communicated by senders (Sperber et al., 2010). To protect themselves from misleading messages, people evaluate both the content and the source of communicated information: if either is found wanting, then the information is rejected. However, these mechanisms have stringent limits: sometimes people would be better off accepting information that flies in the face of their existing beliefs and that is communicated by a source that is not entirely trusted—sometimes others, even others we don't fully trust, know

better than us. This 'trust ceiling' affects both senders and receivers: senders fail to transmit their message, and receivers miss out on potentially valuable information. A solution is for senders to provide reasons supporting the message they want to transmit. Receivers can then evaluate these reasons and, if they are deemed sufficient, change their mind, not because they trust the sender, but because it makes more sense for them to accept her message than to reject it. According to the argumentative theory, reasoning evolved chiefly to enable argumentation.

This hypothesis readily accounts for some features of reasoning described above. First, it is essential that reasoning should be able to reject weak arguments—otherwise manipulation would be altogether too easy, and people would be better off not listening to any arguments. This is what we observe: when it matters, people are not easily swayed by poor arguments.

When it comes to producing arguments, conviction is most likely achieved by finding arguments that support the reasoner's side or go against her interlocutor's. The myside bias is a normal feature of reasoning when it is understood as performing an argumentative function.

However, if the function of reasoning is to convince, one might expect it to produce strong arguments. As we have seen, this is often not the case. We presently offer an explanation for this feature of reasoning that relies on reasoning being used in dialogic contexts—as the argumentative theory would predict. In the next section, we introduce the relevant properties of dialogic contexts in the general case, before exploring in more details the case of argumentation.

Repair in communication

Even though an interlocutor can understand what a speaker means without accepting the message, conversational contexts entail a very high overlap of interests: the speaker wants the interlocutor to understand what she means, and the interlocutor wants to understand what the speaker means. As a result, the burden of making communication as efficient as possible does not fall only on the speaker, but is shared by the interlocutor, a division of labor that has been studied in linguistics (e.g. Clark & Wilkes-Gibbs, 1986; Sacks, Schegloff, & Jefferson, 1974; Schegloff, Jefferson, & Sacks, 1977; Schegloff & Sacks, 1973)..

In the following example the first speaker, *A*, wants the interlocutor, *B*, to understand that he's referring to the Ford family (from Sacks & Schegloff, 1979, p. 19; cited in Levinson, 2006).

> *A*: … well I was the only one other than the uhm tch Fords?,
> Uh Mrs Holmes Ford?
> You know uh the the cellist?
> [
> *B*: Oh yes. She's she's the cellist
> *A*: Yes well she and ……

A starts by simply saying "the Fords," but he does not stop there. He then points out one member of the family in particular ("Mrs Holmes Ford") before specifying her occupation ("the cellist"). This repair was likely initiated by the lack of expected positive feedback following the first attempt to refer to the Ford family. Simply by failing to communicate that she understood who the Fords are, *B* made it clear that she required more information to understand the referent of *A*'s utterance (see, e.g. Goodwin, 1981).

Whether they are 'self-repairs' (as in the present example) or 'other-repairs,' repairs are ubiquitous in verbal communication. Such repairs can follow genuine failures, for instance when the speaker chooses the wrong word. The present example might seem to reflect a failure as well: failure of the speaker to choose the optimal way to refer to the Ford family from the start. However, as argued by the linguists who have studied these repairs, such interactions should instead be understood as reflecting the efficient working of communication.

For *A* to find the best way to refer to the Ford family, he needs to know how *B* knows the Fords. This information could be easily accessible—if, for instance, *A* knew very well that *B* was good friend with the Fords—in which case *A* would have no trouble referring to them. Or it might be nearly impossible to access: maybe *A* knows neither *B* nor the Fords very well, and has only vague hunches about how much they know each other. In this case, *A* has three possible strategies. The first strategy is to think long and hard about whether *B* knows the Fords or not, dig into his memory and his inferential abilities to make the best possible guess. The second strategy is to provide *B* with an exhaustive list of the information he has about the Fords to maximize the chances that *B* understands who *A* is talking about. The third strategy is the one *A*

chooses: to start with the common way of referring to a family in this context, and proceed to offer more clues to who they are until *B* indicates that she understands.

Although the first two solutions seem superficially more efficient—there might be fewer conversational turns—they are in fact more costly: *A* either has to take time and energy to figure out something that would be trivially revealed in the course of the conversation, or *A* has to make a long speech that might be irrelevant if *B* recognizes the Fords immediately. By contrast, a few conversational turns offer a very economic alternative: what looks like a failure might in fact be the most efficient system given the constraints.

'Repair' in argumentation

How does this logic apply to argumentation? If the argumentative theory of reasoning is correct, argumentation solves a problem that affects both senders and receivers, so that senders have an incentive to communicate the best available reasons for their messages, and receivers have an incentive to understand what these reasons are. As in other forms of communication, the alignment of interests isn't perfect—interlocutors can understand speakers' reasons without accepting them—but it is strong.

As a result, the logic described above also applies to argumentation. Instead of laboring to find the strongest possible argument from the start, interlocutors can make the best of the interactive context and refine their arguments as the exchange unfolds. Indeed argumentation should rely even more on feedback than other forms of communication. Finding good arguments is likely to be harder than finding, say, the best way to refer to someone. Fortunately, the difficulty of the task is mitigated by the richness of the feedback. Instead of a mere indication of understanding or failure of understanding, interlocutors who reject an argument often state their reasons for doing so, offering the speaker an opportunity to understand and address these reasons.

Take the following excerpt from a discussion between three students, on the topic of nuclear power (from Resnick, Salmon, Zeitz, Wathen, & Holowchak, 1993, p. 350):

C4: Well, uh is, is nuclear, I'm against it . . Is nuclear power really cleaner that fossil fuels? I don't think so

A5: You don't think, I think//

B6: In terms of atmospheric pollution I think that . . the waste from nuclear power, I think it's . . much less than fossil fuels . . but the waste that there is of course is quite dangerous//

C7: It's gonna be here for thousands of years, you can't do anything with it. I mean, right now we do not have the technology as//

B8: Acid rain lasts a long time too you know

C9: That's true but if you reduce the emissions of fossil fuels which you can do with, uh, certain technology that we do have right now, um, such as scrubbers and such, you can reduce the acid rain, with the nuclear power you can't do any, I mean nuclear waste you cannot do anything with it except//

B10: bury it

A11: m-hm

C12: bury it and then you're not even sure if its ecologically um . . that the place you bury it is ecologically sound.

B13: I, I think if if enough money is spent it can probably be put in a reasonably safe area

…

Here we can see instances of 'self initiated repair,' for instance at *C7*, when *C* specifies that what he meant was that it is impossible to get rid of nuclear waste for good. A rebuttal to a counter-argument can be seen as a form of 'other initiated repair,' as for example in *C7*/*B8*/*C9*, when *C* addresses *B*'s counter-argument by spelling out his argument in more detail: it's not only that nuclear wastes are long lasting, but that, given current technology, the damage created by other wastes is shorter lived than that of nuclear waste.

C could have tried to anticipate the counter-argument offered by *B*. However, in doing so he would have been likely to think of counter-arguments that *B* would never have though of, or that she wouldn't subscribe to, and to miss the counter-argument she actually offered. In most cases such anticipation has high costs—cognitively—and little benefits—since the interlocutor will give her counter-arguments herself. So why bother?

People's ability to adapt and refine their arguments in the course of a discussion has been observed in various contexts such as discussions of contentious topics (Kuhn & Crowell, 2011; Resnick et al., 1993), of logical tasks (Trognon, Batt, & Laux, 2011; Trognon, 1993), and of

classroom tasks (Anderson, Chinn, Chang, Waggoner, & Yi, 1997). However, on the whole it remains an understudied topic.

The various ways in which arguments can fail to convince[1]

The explanation above is not very fined-grained: it accounts for the overall limited quality of arguments, especially those most studied by psychologists which correspond to what should only be the first turn of a discussion. To better understand what it means that people produce relatively weak arguments, it is useful to look into more details at the various ways in which arguments can fail to convince their intended audience. We suggest that there are two main stages at which this can happen.

The first stage bears on the intrinsic quality of the argument: is it a reason at all to support the conclusion? An argument that is found wanting at this point can be called *defective*. In turn, an argument can be defective in different ways, which can be categorized as *external* and *internal*. When an argument is externally defective, the audience either disagrees with, or simply misses, a premise (often an implicit premise). Here are two examples:

(1) Laura: "You should go see this movie, it's by Stanley Kubrick."
George: "I don't know who that is."

(2) Laura: "You should go see this movie, it's by Stanley Kubrick."
George: "I don't really like his movies."

In (1), the argument fails because Laura didn't anticipate that George would not know the implicit premise ("movies by Stanley Kubrick are worth watching."), in (2) it fails because George disagrees with it.

[1] We are not taking a normative stance here, and making for instance distinction between whether the argument should convince based on its soundness or validity, or whether it should merely 'persuade.' We aim at describing psychological mechanisms, so that conviction is obtained when, or to the extent that, the interlocutor changes her mind. Accordingly, we urge the reader to not think of any normative framework in reading what follows (e.g. when we will introduce 'intrinsic quality,' it will refer only to the way it will be defined here, not to the more general notion of logical validity for instance.

By contrast, an internal failure happens when the argument is inherently flawed, as in (3):

(3) Laura: "You should go see Dr. Strangelove rather than Eyes Wide Shut, it's by Stanley Kubrick."

George: "But Eyes Wide Shut is also by Kubrick!"

In this case nothing can be done to salvage the argument. In a simple logical or mathematical task, all the arguments for any wrong answer must fail internally.

Even an argument that is not found to be defective can fail to convince at the next stage of argument evaluation because it is *too weak*. For instance:

(4) Laura: "You should go see Eyes Wide Shut, it's by Stanley Kubrick."

George: "I love Kubrick, but I really hate Tom Cruise, so I think I'll pass."

Here Laura's argument is accepted by George, but it is not sufficient to convince him. The argument is simply not strong enough to change his mind. Even though we will call this a failure here for simplicity, arguments that are found to be weak range from the too weak to have any effect to the nearly strong enough to tip the scales. In the latter case, adding even a relatively weak argument might suffice, so that even though the initial argument failed to completely convince the interlocutor, it will have played the major role when she eventually changes her mind.

With the exception of the internal failures, all the other types of failures reflect a lack of perspective taking: the speaker fails to properly anticipate that the interlocutor does not hold a given belief, or has a stronger belief in the conclusion than anticipated. These failures are related to more general, and well studied, failures of perspective taking known as curse of knowledge (Birch & Bloom, 2007), false consensus effect (Krueger & Clement, 1994), or simply egocentrism (Nickerson, 1999)

As argued above, these failures (again, with the exception of the internal kind) are often not very costly. They do not mean the conversation is over: more arguments can be adduced. In particular, external failures can be fixed by trying to change the interlocutor's mind about the

problematic premise. Here, Laura could inform George that Kubrick is a widely respected director—in (1)—or try to convince George of the value of Kubrick's movies—in (2).

Failed argument or successful explanation?

We will now argue that in some cases these failures are only apparent, not real failures: it depends on what the objective of putting forward the argument is. One way in which reasoning can solve 'trust bottlenecks' is by allowing people to provide arguments in order to convince others, as explained above. However, reasoning can also help alleviate problems of trust by enabling people to justify their decisions.

Figuring out why people do the things they do can be fiendishly difficult. When we fail to reconstruct the reasons for a given behavior it will appear irrational. If we based our evaluation of others on the unaided understanding of their behavior, we would often be led to conclude that they are not very competent, and therefore not very reliable or trustworthy. Reasoning can help solve this problem by letting people explain their apparently irrational behaviors. As in the case of argumentation, interlocutors can then evaluate these reasons to see if they are indeed good reasons. This solution is efficient since (a) it is much easier for the person who engaged in a given behavior to provide a reason for it than it is for most observers, and (b) it is easier for the observer to evaluate a reason provided to her than to figure it out on her own.

There are, however, crucial differences in the way rational explanations of behavior (or of thoughts), on the one hand, and arguments on the other ought to be evaluated (for another take on this issue see, e.g. Bex, Budzynska, & Walton, 2012). For an explanation to be good, it has to make sense from the point of view of the speaker. By contrast, an argument has to be good from the point of view of the interlocutor. Accordingly, external failures are not failures anymore—as long as the speaker can provide premises that are simply unknown to the interlocutor. Consider this variation on (3):

(3') Laura: "I think I will go see this movie, it's by Stanley Kubrick."
George: "I don't really like his movies."

Here George' reaction should not be understood as a refutation of Laura's explanation, but simply as a statement of opinion. To the extent that George can easily fill in the implicit

premise—that Laura likes Kubrick—then he should not find the explanation defective, even if he disagrees with the premise.

Similarly, explanations are less likely to be found to be too weak: they do not have to be strong enough to overcome the interlocutor's belief, but simply strong enough to warrant the speaker's belief. Again, consider a variation on the preceding dialogue:

(4') Laura: "I think I will go see Eyes Wide Shut, it's by Stanley Kubrick."
George: "I love Kubrick, but I really hate Tom Cruise."

As long as George doesn't have a reason to think that Laura shares his distaste for Cruise, or that she has a stronger reason to not see this movie, then he should find the explanation sound.

In many psychological experiments, reasoning might be triggered more as a way of justifying the participant's position, making sure that she stands on rational grounds, rather than for trying to convince someone. For instance, in a typical reasoning task, people do not really care if others hold the same beliefs regarding the right answer. If they are motivated to reason, it is more likely to be as a way to ensure that they can provide an explanation for their answer, to show that it is rational. Even when participants are explicitly asked to defend their opinions on, say, public policy, as in Kuhn (ref), they do not actually face someone who disagrees with them and who they would really like to convince. Although argument failures are to be expected even in a genuine argumentative discussion—for the reasons exposed above—people might be more motivated to engage in some perspective taking, and therefore avoid some argument failures, when they really aim at convincing someone of the argument's conclusion rather than of their rationality.

Are others better at detecting internal argument failures?

The production of externally defective arguments and arguments too weak to convince on their own is caused by the costs of perspective taking: the speaker either cannot anticipate, or does not make the effort to anticipate, the beliefs of the interlocutor. This is exactly what one should expect to happen in interactive contexts, when it often makes more sense to let the interlocutor inform the speaker of her beliefs than to force the speaker to anticipate them.

Moreover, if the goal of the speaker is to explain her position rather than to convince, most of these 'failures' are not failures at all.

Internal failures are not so easily explained. They reflect ignorance about the world (rather than about the interlocutor's beliefs), failures of inference or failures of memory. For instance, in most reasoning tasks people provide arguments that are internally invalid and they fail to make the inferences that would enable them to realize this. Crucially, arguments that fail internally are also poor explanations: although they show that the speaker had *a* reason for her position or behavior, they reveal that this was a poor reason even from the speaker's perspective.

Even though internal failures cannot be explained as part of a well-functioning division of cognitive labor between speaker and interlocutor, they would not have to be especially mysterious if it wasn't for one of their features. After all, every cognitive system is bound to make some mistakes, as it does not have infinite resources. What makes internal argument failures interesting, however, is the asymmetry mentioned above: people seem to be better at spotting such failures in others than in themselves. We suggest that there are two types of explanations. The first is simply of a difference in background beliefs or in their accessibility between the speaker and the interlocutor. In example (3), George might have more knowledge about Kubrick, or he might have just been thinking about who the director of Eyes Wide Shut is. There is no reason, however, that interlocutors should, on average, be more likely to have access to the relevant beliefs than speakers. What explains the asymmetry is that when the speaker accesses the relevant beliefs, she does not produce the argument at all, so there is no observable behavior. By contrast when the interlocutor does, then we can observe the defective argument being corrected.

The second explanation is more interesting. A speaker produces an argument that is, in fact, internally defective. At first, the interlocutor might simply find it too weak to change his mind, but not defective. He would then likely engage in a search for counter-arguments in order either to justify not changing his mind or to convince the speaker to change her mind (or both). In the process, he might find arguments that support his point of view without attacking the speaker's initial argument—as in (4) for instance. But he might also find arguments that specifically target the speaker's initial argument. Such arguments are likely to reveal the defect in the initial argument—as in (3) for example. In this case, the apparent difference in the way speakers and interlocutors evaluate arguments—that the speaker found the argument good

enough to produce while the interlocutor found it defective—does not reflect a difference in evaluation *stricto sensu*, but a difference in evaluation stemming from the production of a counter-argument by the interlocutor. The fact that the interlocutor is more likely to find such a counter-argument is a simple consequence of the myside bias.

Conclusion

Researchers who have studied argument production generally agree that quality is not very high: people routinely produce arguments that are weak or easily countered. This is a problem for the classical theory of reasoning: if reasoning's task was to improve individual cognition, it should make sure we have good reasons for our beliefs or decisions. But this also seems to be an issue for the argumentative theory of reasoning: if the function of reasoning is to convince, then wouldn't it be better to produce strong, convincing arguments?

We have argued that, counter-intuitively, not aiming at very strong arguments is the best strategy for a device working in an interactive, cooperative context. Finding arguments that will appeal to a particular interlocutor entails having a solid grasp on the interlocutor's beliefs, making it an arduous, cognitively costly task. Instead of trying to anticipate the interlocutor's belief, it is possible to start an argumentative discussion by offering an argument that passes some minimal threshold of quality and wait for the interlocutor's feedback. If the argument doesn't convince the interlocutor, he will often provide the speaker with an explanation. This enables the speaker to mend her argument, to adjust it to the interlocutor's relevant beliefs.

In this perspective, most argument failures are better seen as steps in a normal process of interaction. Moreover, if the goal of the reasoner is to justify her behavior rather than to convince the interlocutor of a given conclusion, then most argument failures aren't even failures to begin with: they are perfectly acceptable explanations.

The only exception are internally defective arguments, arguments that contradict beliefs the speaker ought to have considered. These arguments are not only unconvincing, but they also make for poor explanations. Particularly puzzling is the asymmetry in the evaluation of internally defective arguments: why would the interlocutor be in a better position to spot such failures than the speaker, given that the argument clashes with the speaker's beliefs? We suggested that interlocutors might not, at first, be more likely to spot the defect in the argument, but that in the

process of looking for a counter-argument, they might find one that reveals the defect in the argument. The search for counter-argument is guided by the myside bias; therefore the argumentative theory can also account for the asymmetry in the evaluation of internally defective arguments.

Although the study of argumentative discussion, with the interactions they entail, is fraught with methodological difficulties, it is the best place to reach a genuine understanding of reasoning's strengths and (supposed) failures.

Acknowledgments

We would like to thank Steve Oswald for his very useful feedback.

References

Anderson, R. C., Chinn, C., Chang, J., Waggoner, M., & Yi, H. (1997). On the logical integrity of children's arguments. *Cognition and Instruction*, *15*(2), 135–167.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning*, *1*, 221–235.

Bex, F., Budzynska, K., & Walton, D. (2012). Argument and Explanation in the Context of Dialogue. In T. Roth-Berghofer, D. B. Leake, & J. Cassens (Eds.), *Proceedings of the 7th International Workshop on Explanation-aware Computing* (pp. 6–10).

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382–386.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One*, *6*(1), e15954.

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. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, *67*(4), 596.

Kuhn, D. (1991). *The Skills of Arguments*. Cambridge: Cambridge University Press.

Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological Science*, *22*(4), 545.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*, 177–189.

Levinson, S. C. (2006). On the human "interaction engine." *Roots of Human Sociality: Culture, Cognition and Human Interaction. Berg, Oxford.*

Mercier, H., & Sperber, D. (in prep). *The Argumentative Theory*. Cambridge: Harvard University Press.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning*, *4*(3), 231–248.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology*, *2*, 175–220.

Nickerson, R. S. (1999). How we know-and sometimes misjudge-what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*, 737–759.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, N.J: Prentice–Hall.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, *77*, 562–571.

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. Voss, D. Perkins, & J. Segal (Eds.), *Informal Reasoning and Education* (pp. 83–105). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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* (pp. 323–390). Boston: McGraw-Hill.

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.

Sacks, H., & Schegloff, E. A. (1979). Two preferences in the organization of reference to persons in conversation and their interaction. In G. Psathas (Ed.), *Everyday Language: Studies in Ethnomethodology* (pp. 15–21). New York: Irvington.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 696–735.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 361–382.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica*, *8*(4), 289–327.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, *34*(3), 619–632.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, *25*(4), 359–393.

Stanovich, K. E. (2004). *The Robot's Rebellion*. Chicago: Chicago University Press.

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.

Trognon, A., Batt, M., & Laux, J. (2011). Why is dialogical solving of a logical problem more effective than individual solving?: A formal and experimental study of an abstract version of Wason's task. *Language & Dialogue*, *1*(1).

Trouche, E., Sander, E., & Mercier, H. (submitted). Arguments, more than confidence, explain the good performance of groups in intellective tasks.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology: I* (pp. 106–137). Harmandsworth, England: Penguin.

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2015) 1–15

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12303

The Selective Laziness of Reasoning

Emmanuel Trouche,[a] Petter Johansson,[b,c] Lars Hall,[b] Hugo Mercier[d]

[a]*CNRS, Laboratory for Language, Brain and Cognition*
[b]*Cognitive Science, Lund University*
[c]*Swedish Collegium for Advanced Study, Uppsala University*
[d]*Center for Cognitive Sciences, University of Neuchâtel*

Received 7 October 2014; received in revised form 14 July 2015; accepted 16 July 2015

Abstract

Reasoning research suggests that people use more stringent criteria when they evaluate others' arguments than when they produce arguments themselves. To demonstrate this "selective laziness," we used a choice blindness manipulation. In two experiments, participants had to produce a series of arguments in response to reasoning problems, and they were then asked to evaluate other people's arguments about the same problems. Unknown to the participants, in one of the trials, they were presented with their own argument as if it was someone else's. Among those participants who accepted the manipulation and thus thought they were evaluating someone else's argument, more than half (56% and 58%) rejected the arguments that were in fact their own. Moreover, participants were more likely to reject their own arguments for invalid than for valid answers. This demonstrates that people are more critical of other people's arguments than of their own, without being overly critical: They are better able to tell valid from invalid arguments when the arguments are someone else's rather than their own.

Keywords: Reasoning; Argumentation; Choice blindness; Belief bias

1. Introduction

The way people produce arguments is doubly problematic. First, they mostly find arguments for their own side. Second, these arguments tend to be relatively weak. The first trait of argument production—the confirmation bias or myside bias—has been the topic of much attention (see, e.g., Nickerson, 1998). The later has been comparatively neglected, but is well supported by the existing evidence. When asked to justify their points of view, many participants can only generate arguments that make "superficial

Correspondence should be sent to Hugo Mercier, Centre de Sciences Cognitives, Université de Neuchâtel, Espace Louis-Agassiz 1, Neuchâtel 2000, Switzerland. E-mail: hugo.mercier@unine.ch

sense" (Perkins, 1985, p. 568), and they fail to offer genuine evidence (Kuhn, 1991). Similar results have been observed in social psychology (Nisbett & Ross, 1980) and in the study of formal reasoning (Evans, 2002). When people face simple problems ranging from the Wason selection task (Wason, 1966) to the Cognitive Reflection Test (Frederick, 2005), they typically start with a wrong intuition, which the subsequent reasoning fails to correct in most cases. This happens not only because people mostly look for arguments supporting their intuition (see Ball, Lucas, Miles, & Gale, 2003), but also because they are satisfied with the arguments they find—arguments that must be flawed given that they support a logically or mathematically invalid answer. Summarizing the perspective of dual process theories, Kahneman (2011) explains this poor performance of reasoning by the fact that "System 2 is sometimes busy, and often lazy" (p. 81): Reasoners do not make the effort that would be required to produce better arguments (see also, e.g., Evans, 2008).

This laziness, however, does not seem to apply to all arguments. When people evaluate other people's arguments—in particular, if they disagree with their conclusion—they appear to be more careful, and to mostly accept strong arguments. This result has been observed in research on persuasion and attitude change (for a review, see Petty & Wegener, 1998), and in Bayesian studies of argumentation (Hahn & Oaksford, 2007). Sound argument evaluation skills are also indicated by the fact that participants are convinced by arguments supporting the valid answer to reasoning problems such as those mentioned above (for the Wason selection task, see Moshman & Geil, 1998; for the CRT, see Trouche, Sander, & Mercier, 2014; and, more generally, Laughlin, 2011).

When it comes to evaluating others' arguments, the evaluation is most likely to be thorough when participants disagree with the argument's conclusion. When they agree with an argument's conclusion, not only are participants more likely to find the argument valid, but they also discriminate less between valid and invalid arguments, showing a relaxation of their evaluative criteria (Evans, Barston, & Pollard, 1983). Given that when participants produce arguments, they agree with the argument's conclusion, a more general way to frame the asymmetry between argument production and argument evaluation is as follows. When people agree with an argument's conclusion, they tend to evaluate it only superficially—this includes others' arguments whose conclusion one agrees with or arguments one produces. When people disagree with an argument's conclusion, they tend to evaluate it more thoroughly. Reasoning would thus only be *selectively* lazy.

The asymmetry that has the greatest ecological validity is that between the production of arguments and the evaluation of arguments whose conclusion one disagrees with—this is what happens in a standard exchange of arguments in which two or more people try to convince each other of their respective viewpoints. However, this asymmetry has only been indirectly demonstrated, from comparisons of disparate studies, and it is confounded by the fact that argument quality varies between different contexts and interlocutors. A convincing demonstration of this asymmetry would instead involve participants evaluating *their own arguments as if they were someone else's*. We would then expect that the participants would reject many of the arguments they deemed good enough to produce, if they thought the arguments came from someone else and they disagreed with their

conclusion. Moreover, they should be better at discriminating between their own good and bad arguments when they think they are someone else's and they disagree with their conclusion.

To test this prediction, we relied on the choice blindness paradigm, in which participants are led to believe that they have provided a given answer when in fact they answered something else. For example, in Hall, Johansson, and Strandberg (2012), the participants rated to what extent they agreed with moral issues, such as "If an action might harm the innocent, it is morally *reprehensible* to perform it." Using a sleight of hand, the participants' answers were at times reversed: If they had indicated that they agreed with the preceding statement, their answer now read that they agreed with an opposite statement (i.e. "... it is morally *permissible*..."). Participants were then asked to defend their positions, so that they would sometimes be asked to defend a moral position that was the opposite of their originally stated position. Not only did more than half of the participants often miss the switch, but they also gave coherent and detailed arguments supporting the opposite of their original opinion.

This general finding has been replicated in a number of different contexts and domains. Choice blindness has been demonstrated for attractiveness of faces (Johansson, Hall, Sikström, & Olsson, 2005; Johansson, Hall, Sikström, Tärning, & Lind, 2006; Johansson, Hall, Tärning, Sikström, & Chater, 2014), moral and political choices (Hall et al., 2012, 2013), and financial decision making (McLaughlin & Somerville, 2013). In addition, choice blindness has been demonstrated for taste and smell (Hall, Johansson, Tärning, Sikström, & Deutgen, 2010), for tactile stimuli (Steenfeldt-Kristensen & Thornton, 2013), and for auditory stimuli (Lind, Hall, Breidegard, Balkenius, & Johansson, 2014; Sauerland, Sagana, & Otgaar, 2013).

In the present case, we use a choice blindness manipulation in a reasoning task to make people believe that an answer and an argument they previously provided had been generated by another participant. The main prediction of the selective laziness account is that participants would reject many of the arguments they previously made, in particular bad arguments. By contrast, they should be more likely to accept their own good arguments.

2. Experiment 1

2.1. Method

2.1.1. Participants

We recruited 237 participants (100 females, M_{age} = 34.2, SD = 12.0) residing in the United States through the Amazon Mechanical Turk website. The total N was reached in two sessions: first an N of 160 and then an N of 77. We estimated the N for the first session based on the low detection rates (see below for an explanation of detection rates) obtained in previous choice blindness manipulations which would have allowed us to retain most of the participants for the comparison between manipulated and non-manipulated problems. However, the relatively higher rate of detection in the current experiment

allowed us to keep fewer participants than expected in the manipulated condition. The second session was conducted to approach the *N* initially aimed at for the manipulated condition. The experiments took about 10 min to complete, and we paid the participants standard rates for participation ($0.7).

2.1.2. Procedure and materials

The experiment consisted of two phases. In Phase 1, we presented the participants with five enthymematic syllogisms—syllogisms with an implicit premise—in succession (for an example syllogism see Fig. 1; all syllogisms can be found in the section "Materials" in the Supporting Information). For each syllogism, we asked the participants to choose which of five alternatives they thought was the valid answer and to explain why they gave their chosen answer (see Fig. 1).

At the start of Phase 2, which took place right after Phase 1, we told the participants that all five problems would be presented again, accompanied by the answer and explanation from another participant. The complete instruction read:

> We will now proceed to the second phase of the experiment. For each of the five problems, we will give you the answer given by another participant along with the explanation they gave. Each answer was provided by a different participant. You will be able to change your answer in light of this information if you wish.

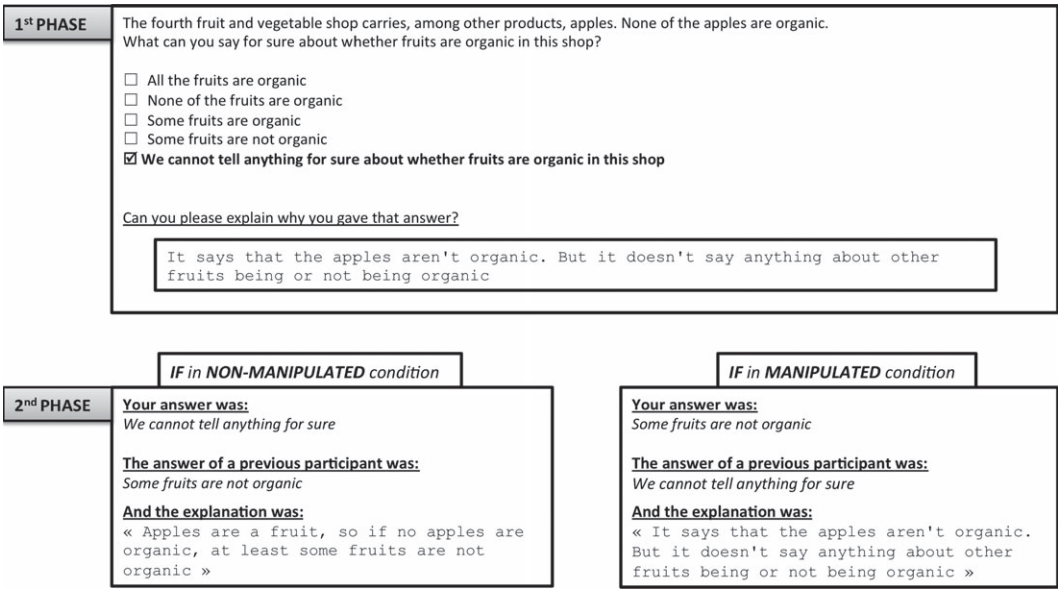


Fig. 1. Example of syllogism used in the experiment, shown both in the Manipulated and Non-Manipulated alternatives.

When the syllogisms were presented the second time, the participants were reminded of their own previous answer and provided with what was presented as someone else's answer and argument. They were told that they could change their answer in light of this information. The answer they had to evaluate was either the valid one (if the participants had previously given an invalid answer) or the most common invalid answer (if they had previously given the valid answer). The arguments presented for each answer were the same for all participants, constructed to be plausible explanations for making that particular (valid or invalid) choice (all the arguments used can be found in the Supporting Information).

However, for one of the syllogisms (the manipulated syllogism), instead of being truthfully reminded of their previous answer, participants were told that they had given an answer different from the one they had given: either the valid answer (if they had answered invalidly) or the most common invalid answer (if they had answered validly). Their own previous answer, and the argument that justified it, were presented as if they were those given by another participant. The external features of the presentation were strictly identical to those of the other four syllogisms (see Fig. 1 for an example of both conditions).

There were two different conditions, in which a different syllogism was manipulated. For half of the participants ($N = 119$), the EA3 syllogism was manipulated, and for the other half ($N = 118$) the EA4 was manipulated (see Supporting Information for the list of all syllogisms). This was to ensure that the outcome was not dependent on the properties of a particular syllogism. The Manipulated trials appeared on the fourth trial for all participants.

At the end of the experiment, we asked the participants a series of questions to determine if they had detected the manipulation of their answer. The first question simply asked what they thought of the experiment, the second if they thought anything was strange with the experiment, and in the third they were then presented with an actual description of the manipulation and asked if they had noticed this:

> During the second phase, in one or more problem your previous answers had been changed to another answer, and the "someone else's answer and argument" were actually the ones you had previously given. Did you notice this?

If they answered yes, we asked them how many times they had noticed this, followed by a question asking them to specify which of the five problems they thought had been altered. Finally, we asked the participants about their gender, age, and level of education.

2.1.3. Results

2.1.3.1. Detection rates: When the manipulation procedure was described in the debriefing, 53% of the participants indicated that they had noticed that their answers had been altered. Of these 53%, only 32% were able to retrospectively identify which of the syllogisms had actually been manipulated. This suggests that the 53% measure is conservative,

in that it includes participants who might not in fact have detected the manipulation. By contrast, it seems unlikely that the participants who said they had not detected the manipulation in fact had done so.

Using the conservative self-reported 53% as a criterion of detection (Detectors), we can ensure that the participants classified as Non-Detectors were in fact unaware of the manipulation, and thus treated their own argument as if it was someone else's (see Hall et al., 2012, 2013; Johansson et al., 2005 for further discussions of participant awareness and detection criteria in choice blindness experiments). The two target syllogisms did not differ in detection rates (EA3: 54%; EA4: 53%; Fisher's exact, $p = .90$, $OR = 0.95$), and they are therefore combined in the remainder of the analysis.

The participants who had given the right answer during the first phase were more likely to detect the manipulation compared to those that had given the wrong answer ($M = 64\%$ and 41%; Fisher's exact, $p < .001$, $OR = 2.5$). In Phase 2, three participants made a decision that was difficult to categorize: They neither stuck with the answer attributed to them nor accepted the argument, choosing instead a different invalid answer. Because their behavior was not easily interpretable in terms of sticking to one's answer versus accepting the arguments, which is the most relevant analysis for Phase 2, they were removed altogether from the analyses (including this statistic).

2.1.3.2. Phase 1 results: The mean score over five problems for all participants, counting 1 for a valid answer and 0 for an invalid answer, was 2.9 ($SD = 1.3$). Success rates for the syllogisms that were manipulated in Phase 2 were 60% for EA3 and 46% for EA4. It was, on average, 41% for Non-Detectors. For all the results that follow, we focus on the most relevant subset: the results from Non-Detectors on Manipulated problems. All the other results can be found in the Supporting Information.

2.1.3.3. Phase 2 results, rates of valid answers and comparison with Phase 1: Sixty-two percent of Non-Detectors gave a valid answer at Phase 2 on the Manipulated problem, a significant improvement over Phase 1 (Phase 1: 41% correct, exact McNemar's test, $\chi^2(1) = 7.93$, $p = .005$). The improvement in Phase 2 was also found for Non-Detectors as well as Detectors on the Non-Manipulated problems (see Supporting Information for details).

2.1.3.4. Phase 2 results, reaction to the arguments: Participants could react to the presentation of the arguments in three ways: they could keep the answer they had given in Phase 1 (or that had been attributed to them in the Manipulated trials), they could accept the answer supported by the argument (someone else's argument in the Non-Manipulated trials, their own argument in the Manipulated trials), or they could pick some other, new answer. As mentioned above, the three participants belonging to the last category have been removed from all analyses.

In Manipulated problems, Non-Detectors participants rejected what was in fact their own argument on 56% of the trials ([18 + 42]/[18 + 26 + 42 + 22] see Fig. 2). Participants who had given an invalid answer in Phase 1, and who were therefore presented

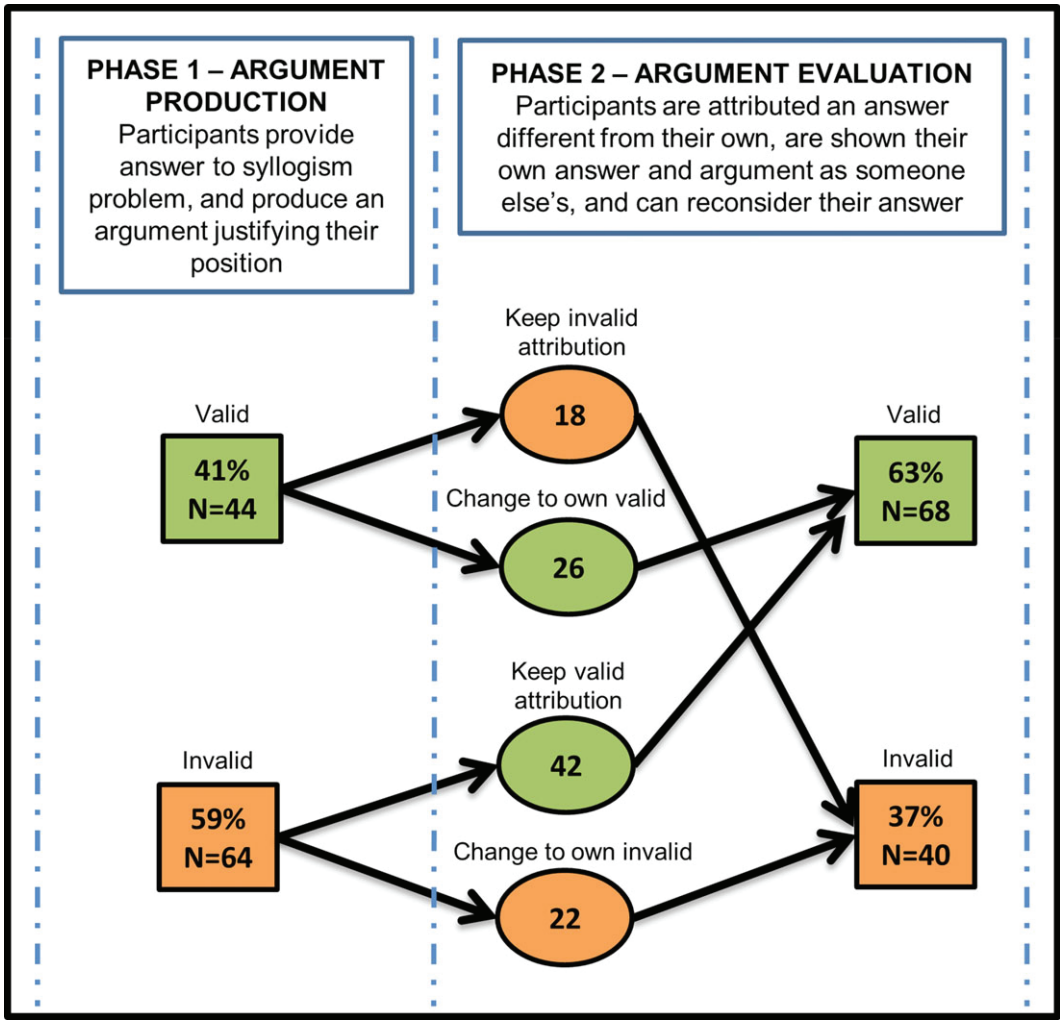


Fig. 2. Results for Non−Detectors on the Manipulated syllogism in Experiment 1. Boxes show percentages of valid and invalid answers at the end of each phase of the experiment. Arrows and ovals show the transition to valid and invalid answers as a consequence of the instructions in each phase. In Phase 1, participants provided the answer to a syllogism and provided an argument for their answer. In Phase 2, participants were attributed an answer different from their own and were confronted with their own Phase 1 answer and argument as if they were someone else's. They could then revise their attributed answer in light of this argument.

with their own argument for this invalid answer were more likely to reject the argument (42/[42 + 22] = 66%) than those who had given the valid answer in Phase 1, and who were therefore presented with their own argument for this valid answer (18/[18 + 26] = 41%) (Fisher's exact test, $p = .02$, $OR = 2.7$). For most of the Non-Manipulated syllogisms, both Non-Detectors and Detectors were also more likely to reject arguments for invalid than for valid answers (see Supporting Information).

2.1.4. Discussion

The goal of Experiment 1 was to compare the difference between how participants treat an argument when they produce it themselves and when they evaluate it as if it was someone else's. In Phase 1, participants were asked to solve reasoning problems and to provide arguments for their answer. If they evaluate their own arguments critically, they should realize that arguments for invalid answers are flawed and adopt valid answers. Thus, invalid answers reflect a poor evaluation of one's own arguments. In Phase 2, in which participants were asked to evaluate others' arguments, one problem was manipulated so that participants were in fact evaluating their own argument. Among the 47% who did not detect the manipulation, 56% rejected their own argument, choosing instead to stick to the answer that had been attributed to them. Moreover, these participants (Non-Detectors) were more likely to accept their own argument for the valid than for an invalid answer.

These results shows that people are more critical of their own arguments when they think they are someone else's, since they rejected over half of their own arguments when they thought that they were someone else's. It also shows that participants can discriminate strong from weak arguments when they think they are someone else's. However, a limitation of Experiment 1 is that it does not provide a good measure of how critical people are toward their own arguments when they produce them. Phase 1 performance is a mix of two factors. First, participants' initial intuitions (or initial models, see, e.g., Johnson-Laird & Bara, 1984), which might guide them toward the valid answer (see, e.g., Sperber, Cara, & Girotto, 1995). Second, participants' reasoning about this initial intuition. It is thus possible that reasoning played no positive role in Phase 1, even when participants provided the valid answer. Current "default-interventionist" models fit with this description of the participants' behavior in Phase 1, as they emphasize that reasoning often does not intervene to modify the intuitive, "default" answer (Evans, 2006; Kahneman, 2011; Stanovich, 2011).

To better examine the role of reasoning in the evaluation of one's own arguments, in Experiment 2 we introduced a Phase 0 in which participants had to provide a quick and intuitive answer to the same problems. In Phase 1, they were asked to justify this answer, and they could change their answer if they wished. This manipulation was similar to the "two responses" paradigm that has been previously used to study metacognitive monitoring (see, e.g., Thompson, Turner, & Pennycook, 2011; it should be noted that in the present experiment no metacognitive questions were asked).

Phase 2 was then identical to the Phase 2 of Experiment 1. This second experiment also addresses a potential concern with Experiment 1: The difference in performance between the production and evaluation of arguments might simply reflect the fact that people were confronted with the same problems for a second time when they were asked to evaluate arguments. As a result, they might simply have learned how to better solve this type of problems. In Experiment 2, the second presentation of the same problems occurs as people are asked to produce arguments, so it will be possible to tell if the mere repetition of the problems leads to improved performance.

Another possible concern is that the participants in the experiment might be afraid of not receiving their payment if they reported the manipulation. Experiment 2, tested for this possibility by introducing, for half the participants, a disclaimer prior to the debriefing reassuring the participants that they would get paid whatever they replied to the debriefing questions.

3. Experiment 2

3.1. Method

3.1.1. Participants

We recruited 174 participants (61 females, M_{age} = 35.5, SD = 12.0) residing in the United States through the Amazon Mechanical Turk website. The experiments took about 10 min to complete, and we paid the participants standard rates for participation ($0.7).

3.1.2. Procedure and materials

The experiment consisted of three phases. In Phase 0, the participants were presented with the five enthymematic syllogisms used in Experiment 1. For each syllogism, the participants were asked to choose which of five alternatives they thought was the valid answer. Participants were asked to provide a "fast, intuitive answer."

In Phase 1, participants were presented with the same problems, reminded of their initial answer, asked to provide an argument for this answer, and offered the possibility to give a new answer. No problem was manipulated in Phase 1.

Phase 2 was identical to the Phase 2 of Experiment 1, using the answers and arguments provided at Phase 1. Given that we had observed no interesting difference between the two syllogisms manipulated in Experiment 1, in this experiment we only manipulated one syllogism, an EA3.

The debriefing phase was identical to that of Experiment 1 with one exception: For half of the participants, it was preceded by a disclaimer reassuring them that they would get paid whatever they replied to the debriefing questions. Finally, we asked the participants about their gender, age, and level of education.

3.1.3. Results

Due to its design, Experiment 2 yielded a rich set of results, most of which do not speak to the point in hand. Here we focus on the results that pertain to the hypotheses laid out above. Other results can be found in the Supporting Information.

3.1.3.1. Detection rates: The debriefing manipulation (i.e., reassuring participants that they would get paid anyway) had no effect on detection rates (detection rate with special debriefing: 48%, normal debriefing: 44%, Fisher exact test p = .76, OR = 1.1); all results are thus presented together. When the manipulation procedure was described in the debriefing, 46% of the participants indicated that they had noticed that their answers had

been altered. Among the 46% of Detectors, only 40% were able to retrospectively identify which of the syllogisms had actually been manipulated. Using the inclusive self-reported 46% as a criterion of detection, we can ensure that the participants classified as Non-Detectors were in fact unaware of the manipulation, and thus treated their own argument as if it was someone else's. The participants who had given the right answer during Phase 1 were more likely to detect the manipulation compared to those that had given the wrong answer (M = 61% and 30%; Fisher's exact, $p < .001$, OR = 3.6). In Phase 2, eight participants made a decision that was difficult to categorize: they neither stuck with the answer attributed to them nor accepted the argument, choosing instead a different invalid answer. For the same reasons as in Experiment 1, they were removed altogether from the analyses (including this statistic).

3.1.3.2. Phase 0 results: The mean score over the five syllogisms for all participants, was 2.9 (SD = 1.19). It was 63% for the syllogism that was manipulated in Phase 2 (53% for Non-Detectors).

3.1.3.3. Phase 1 results: The mean score over the five syllogisms for all participants, counting 1 for a valid answer and 0 for an invalid answer, was 2.9 (SD = 1.30). It was 57% for the syllogism that was manipulated in Phase 2 (43% for Non-Detectors). The effects of participants' attempting to justify their intuitive answers can be broken down into the following categories. The participants who had initially provided a valid answer could either keep this valid answer, or change to adopt an invalid answer (see Fig. 3 for the results of Non-Detectors on the Manipulated problem). The participants who had initially provided an invalid answer could either keep this invalid answer, change to adopt the valid answer, or change to adopt another invalid answer. These results show that participants were more likely than not to keep their intuitive answer (23 changed and 141 did not, Binomial test, $p < .001$) and that they were not more likely to keep their initial answer when it was valid than when it was invalid (13% changed when valid, 17% changed when invalid, Fisher's exact, $p = .49$). The same pattern was observed among Non-Detectors for the Manipulated problem (note that this problem has not been manipulated yet) (19 changed and 67 did not, $p < .001$; 24% changed when valid, 20% changed when invalid, $p = .80$). For all the results that follow, we focus on the most relevant subset: the results from Non-Detectors on Manipulated problems. All the other results can be found in the Supporting Information.

3.1.3.4. Phase 2 results, rates of valid answers and comparison with Phase 1: Sixty-four percent of Non-Detectors gave a valid answer at Phase 2 on the Manipulated problem, a significant improvement over Phase 1 (Phase 1: 43% correct, exact McNemar's test, $\chi^2(1) = 5.78$, $p = .02$). The improvement in Phase 2 was also found for Non-Detectors as well as Detectors on the Non-Manipulated problems (see Supporting Information).

3.1.3.5. Phase 2 results, reaction to the arguments: Participants could react to the presentation of the arguments in three ways: they could keep the answer they had given (or

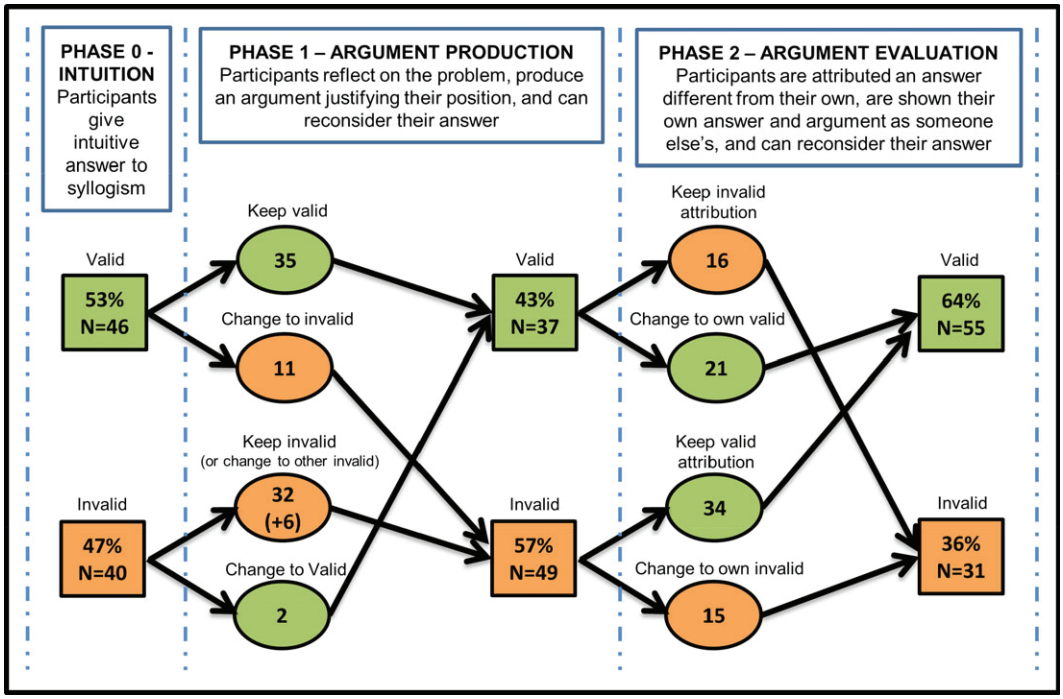


Fig. 3. Results for Non-Detectors on the Manipulated syllogism in Experiment 2. Boxes show percentages of valid and invalid answers at the end of each phase of the experiment. Arrows and ovals show the transition to valid and invalid answers as a consequence of the instructions in each phase. In Phase 0, participants provided an intuitive answer to a syllogism. In Phase 1, participants reflected on the same syllogism and provided an argument for their answer, which they could reconsider in the process. In Phase 2, still on the same syllogism, participants were attributed an answer different from their own, and were confronted with their own Phase 1 answer and argument as if they were someone else's. They could then revise their attributed answer in light of this argument.

that had been attributed to them in the Manipulated trials), they could accept the answer supported by the argument (someone else's in the Non-Manipulated trials, their own argument in the Manipulated trials), or they could pick some other, new answer (see Fig. 3). As mentioned above, the eight participants belonging to the last category have been removed from all analyses.

In Manipulated problems, Non-Detectors rejected what was in fact their own argument on 58% of the trials ([16 + 34]/[16 + 21 + 34 + 15]). Participants who had given an invalid answer in Phase 1, and who were therefore presented with their own argument for this wrong answer were more likely to reject the argument (34/[34 + 15] = 69%) than those who had given the valid answer in Phase 1, and who were therefore presented with their own argument for this valid answer (16/[16 + 21] = 43%) (Fisher's exact test, $p = .02$, $OR = 2.9$). We observed similar patterns for the Non-Manipulated syllogisms, both for Non-Detectors and Detectors (see Supporting Information).

3.1.4. Discussion

The goal of Experiment 2 was to further test the hypothesis tested in Experiment 1, namely, that participants are less critical of the same argument when they produce it than when they evaluate it as if it were someone else's. In particular, Experiment 2 aimed at a better understanding of the effects of argument production. Here, we focus on the Manipulated problem, for which we can compare the effects of argument production (in Phase 1), and the effects of the evaluation of the same argument when participants think it is someone else's (in Phase 2).

In Phase 0, participants provided an intuitive answer to the problem. In Phase 1 they were asked to give an argument to justify their answer, which they could then modify. If participants, during argument production, were critical of their own argument, in Phase 1 they would keep their intuitive answers given in Phase 0 when the answers are valid—which means that valid arguments can be found—and dismiss the intuitive answers when they are invalid—which means that no valid argument can be found.

The results from Phase 1 reveal that participants were not critical of their own arguments as they produced them. Not only did they only reject 5% of the invalid answers, but they did not reject valid answers at a different rate (13%). This result replicates previous results obtained with the "two responses" paradigm with similar problems (Thompson et al., 2011). Reasoning thus mostly provided post-hoc justifications for intuitive answers—a common outcome in this type of task (see, e.g., Evans & Wason, 1976)—and displayed no ability to discriminate between the participants' own valid and invalid answers.

By contrast, when participants thought the same arguments were someone else's, they prove more critical, rejecting 58% of the arguments. More important, reasoning also proved more discriminating, rejecting more arguments for invalid (69%) than for valid answers (43%).

4. Discussion

In two experiments, participants were asked to solve a series of simple reasoning problems, to produce arguments for their answers, and then to evaluate others' arguments about these answers. However, one of the problems was manipulated so that in fact participants were asked to evaluate their own argument as if it was someone else's. Across the two experiments, approximately half of the participants did not detect this manipulation. The participants who did not detect the manipulation thus evaluated an argument they had produced a few minutes before as if it was someone else's.

In the two experiments, participants proved critical of their own arguments when they thought that they were someone else's, rejecting more than half of the arguments. They also proved discriminating: They were more likely to reject their own arguments for invalid answers than their own arguments for valid answers.

Experiment 2 provides a contrast between the performance of reasoning when it evaluates others' arguments and when it produces arguments. In this experiment, participants were first asked to give intuitive answers to the problems, before being asked to produce arguments for these answers. The production of arguments had little effect on the answers, with the vast majority of participants keeping their intuitive answer. Moreover, participants were not more likely to change their invalid answers than their valid answers, so that reasoning did not exert any discrimination.

These experiments provide a very clear demonstration of the selective laziness of reasoning. When reasoning produces arguments, it mostly produces post-hoc justifications for intuitive answers, and it is not particularly critical of one's arguments for invalid answers. By contrast, when reasoning evaluates the very same arguments as if they were someone else's, it proves both critical and discriminating.

The present results are analogous to those observed in the belief bias literature (e.g., Evans et al., 1983). When participants evaluate an argument whose conclusion they agree with, they tend to be neither critical (they accept most arguments) nor discriminating (they are not much more likely to reject invalid than valid arguments). By contrast, when they evaluate argument whose conclusion they disagree with, they tend to be more critical (they reject more arguments) and more discriminating (they are much more likely to reject invalid than valid arguments). The similarity is easily explained by the fact that when reasoning produces arguments for one's position, it is automatically in a situation in which it agrees with the argument's conclusion.

Selective laziness can be interpreted in light of the argumentative theory of reasoning (Mercier & Sperber, 2011). This theory hypothesizes that reasoning is best employed in a dialogical context. In such contexts, opening a discussion with a relatively weak argument is often sensible: It saves the trouble of computing the best way to convince a specific audience, and if the argument proves unconvincing, its flaws can be addressed in the back and forth of argumentation. Indeed, the interlocutor typically provides counter-arguments that help the speaker refine her arguments in appropriate ways (for an extended argument, see Mercier, Bonnier, & Trouche, unpublished data). As a result, the laziness of argument production might not be a flaw but an adaptive feature of reasoning. By contrast, people should properly evaluate other people's arguments, so as not to accept misleading information—hence the selectivity of reasoning's laziness.

Acknowledgments

We thank the Swiss National Science Fund for its support through an Ambizione grant to H.M. P.J. thanks the Bank of Sweden Tercentenary Foundation and The Swedish Research Council (2014-1371). L.H. thanks the Bank of Sweden Tercentenary Foundation (P13-1059:1) and the Wallenberg Network Initiative (WNI). E.T. thanks the Direction Générale de l'Armement (DGA) for its support.

Data availability

All the data are available at this URL: https://sites.google.com/site/hugomercier/Data_ChoiceB_Reasoning.xls?attredirects=0

References

Ball, L. J., Lucas, E. J., Miles, J. N., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *The Quarterly Journal of Experimental Psychology*, *56*(6), 1053–1077.

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. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, *59*, 255–278.

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., & Wason, P. C. (1976). Rationalization in a reasoning task. *British Journal of Psychology*, *67*, 479–486.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS ONE*, *7*(9), e45457.

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, *117*(1), 54–61.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLoS ONE*, *8*(4), e60554.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119.

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition*, *15*(4), 673–692.

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change: You will like this paper better if you (believe you) chose to read it! *Journal of Behavioral Decision Making*, *27*(3), 281–289.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, *16*(1), 1–61.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar Straus & Giroux.

Kuhn, D. (1991). *The skills of arguments*. Cambridge, UK: Cambridge University Press.

Laughlin, P. R. (2011). *Group problem solving*. Princeton, NJ: Princeton University Press.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Speakers' acceptance of real-time speech exchange indicates that we use auditory feedback to specify the meaning of what we say. *Psychological Science*, *25*(6), 1198–1205.

McLaughlin, O., & Somerville, J. (2013). Choice blindness in financial decision making. *Judgment and Decision Making*, *8*(5), 561–572.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning*, *4*(3), 231–248.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology*, *2*, 175–220.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, *77*, 562–571.

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* (pp. 323–390). Boston: McGraw-Hill.

Sauerland, M., Sagana, A., & Otgaar, H. (2013). Theoretical and legal issues related to choice blindness for voices. *Legal and Criminological Psychology*, *18*(2), 371–381.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, *57*, 31–95.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.

Steenfeldt-Kristensen, C., & Thornton, I. M. (2013). Haptic choice blindness. *I-Perception*, *4*(3), 207.

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology: I* (pp. 106–137). Harmandsworth, UK: Penguin.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Materials and results from Experiments 1 and 2

5.4 Le raisonnement comme responsable des échecs individuels

Nous avons à présent tous les éléments pour expliquer à la fois les erreurs massives et la sur-confiance lors de la résolution de problèmes incongruents en solitaire. Non seulement le biais vers son côté tend à produire des raisons pour nos intuitions initiales mais, en plus, la paresse sélective de notre raisonnement nous fait évaluer ces raisons comme parfaitement satisfaisantes. Ces deux tendances majeures du raisonnement sont des biais si l'on considère que la fonction du raisonnement est de corriger nos intuitions trompeuses. Dans le cadre de dialogues argumentatifs en revanche, ce sont des traits parfaitement adaptés pour convaincre à moindre coût.

Situons à présent la théorie argumentative du raisonnement par rapport à la classification des explications possibles proposée par De Neys et Bonnefon (2013). Laissons tout d'abord de côté l'explication en termes d'échec de stockage et supposons que, au moins pour les problèmes du bonbon et de la baguette et de Paul et Linda, les sujets pourraient accéder à la réponse.

Nous avons vu que, dans le cas du bonbon et de la baguette, les études de De Neys et collaborateurs ont mis en évidence que, pour les sujets qui donnent la réponse intuitive et incorrecte, leur sentiment de confiance indique une certaine détection de leur erreur. Cela nous pousse-t-il forcément à donner une explication en termes d'échec d'inhibition ? Pas nécessairement. En effet, si l'on considère que le raisonnement n'essaie même pas de corriger nos intuitions mais au contraire cherche des raisons pour les justifier, l'explication à apporter est quasiment inversée : il faut expliquer comment, avec ce mécanisme, certains sujets arrivent tout même, après réflexion, à corriger leurs intuitions. Les données de De Neys et collaborateur montrant que les sujets ont une confiance en leur réponse moins forte dans la version congruente ou incongruente du bonbon et de la baguette peut s'expliquer de notre point de vue par deux facteurs. Premièrement, ce problème fait partie des tâches où des indices de bas niveau peuvent donner un sentiment de doute. Deuxièmement, lorsque les processus de justifications sont enclenchés, des indices sur la qualité de ces raisons peuvent également participer au sentiment de confiance final. On peut penser, par exemple, que lorsque les sujets produisent des

arguments dans la version congruente, ce processus soit ressenti comme plus fluide que lorsqu'il justifie la version incongruente. Une récente étude de Johnson, Tubau et De Neys (2016) a montré que, même lorsque les sujets n'ont pas le temps de raisonner (de produire des justifications), ceux qui donnent la réponse intuitive au problème incongruent affichent une confiance plus basse que ceux qui donnent la réponse intuitive au problème congruent. Si l'on ajoute également nos observations de l'augmentation du niveau de confiance lorsque les sujets ont le temps de raisonner, que ce soit pour ceux qui donnent la bonne réponse ou la mauvaise au problème incongurent, on peut penser que, malgré le fait que le sujets semblent se douter qu'ils n'ont pas lu l'énoncé correctement, la plupart d'entre eux produisent des raisons pour leur intuition initiale et s'en satisfont, affichant une confiance finale malgré tout très élevée. Dans notre expérience avec 226 étudiants de Licence, la confiance des sujets donnant la mauvaise réponse augmente cependant moins vite que celle des sujets ayant la bonne réponse. Cela vient-il d'un sentiment de doute durant la production de justification, du sentiment de doute initial ou des deux ? Ce n'est pas clair, d'autant que la comparaison se fait, dans notre cas, avec des sujets qui, d'une part ont la bonne réponse après avoir eu la mauvaise intuition, d'autre part ont forcément de meilleures raisons à produire puisqu'ils sont du côté de la vérité mathématiques. Quoiqu'il en soit, sans nier le fait que les sujets détectent précocement qu'ils sont dans l'erreur, cela ne les empêche pas de voir leur confiance augmenter au cours du temps. Plutôt qu'une explication en termes d'échec de l'inhibition, nous refusons l'idée même d'échec, si ce n'est celui, pour certains sujets, à justifier leur réponse, ou plutôt, à se rendre compte, au cours de leur justification, qu'ils ont mal lu l'énoncé.

Prenons à présent le cas du problème de Paul et Linda. Comme nous l'avons vu, la confiance des sujets qui donnent la réponse intuitive et de ceux qui donnent la bonne réponse est aussi élevée. Cela reste à montrer mais je suggère qu'il n'y a aucun indice, dans l'énoncé, exploitable par des processus de bas niveaux. Le contraste entre les deux tâches est relativement intéressant à discuter. On peut en effet penser que le fait que les sujets donnant la réponse intuitive finissent aussi confiants que les sujets donnant la bonne réponse vienne du manque d'indice d'erreur dans l'énoncé. Dans ce sens,si l'on faisait résoudre la tâche dans une version congruente à des sujets et incongruente à d'autres, tous

avec une mémoire de travail surchargée, nous prédisons que, d'une part personne ne donnerait la bonne réponse au problème incongruent, d'autre part la confiance en la réponse intuitive ne devrait pas différer entre les versions congruentes et incongruentes. Remarquons cependant que l'expérience ne pourra sans doute pas fonctionner si la mémoire de travail est trop surchargée. En effet, dans la tâche de Paul et Linda, la réponse 'intuitive' n'est pas aussi intuitive que dans le cas du bonbon et de la baguette. La réponse erronée « on ne peut pas savoir » ne saute pas autant aux yeux que la réponse « 0.10€ » dans le problème du bonbon et la baguette. Quoiqu'il en soit, notre explication des échecs dans le cas de la tâche de Paul et Linda est plus proche d'une explication en termes d'échec de détection plutôt qu'en termes d'échec d'inhibition. Les sujets dans l'erreur ne détecteraient, dans ce dernier cas, même pas « qu'il y a un problème ». Cela dit de notre point de vue, rappelons-le, l'explication ne se fait pas en termes d'échec puisque le raisonnement fait l'inverse qu'essayer de corriger nos intuitions. La différence en termes de sentiment de confiance entre les deux tâches pourrait cependant s'expliquer par le manque d'indice d'erreur dans l'énoncé, la tâche de Paul et Linda ne donnant, d'après nous, lieu à aucun signal que « quelque chose ne s'est pas passé comme prévu ».

Pour conclure ce chapitre, la théorie argumentative propose bien plus qu'une norme alternative. Elle propose des explications pour les échecs des sujets à résoudre des tâches incongruentes en solitaire à partir de la remise en question de la fonction du raisonnement. Cette fonction serait argumentative et non correctrice. Elle se traduit, non seulement par une tendance à produire des arguments pour nos intuitions, mais également par une évaluation superficielle de nos arguments. Ainsi le raisonnement, non seulement ne corrige souvent pas nos intuitions, mais nous rend confiant dans nos erreurs.

Jusque-là le portrait psychologique du raisonnement est plutôt noir. Y a-t-il des moyens d'éviter les erreurs et la sur-confiance ? L'échange de raisons avec les autres en est un. Cependant, nous allons le voir, le raisonnement en groupe n'est pas toujours synonyme d'amélioration des performances.

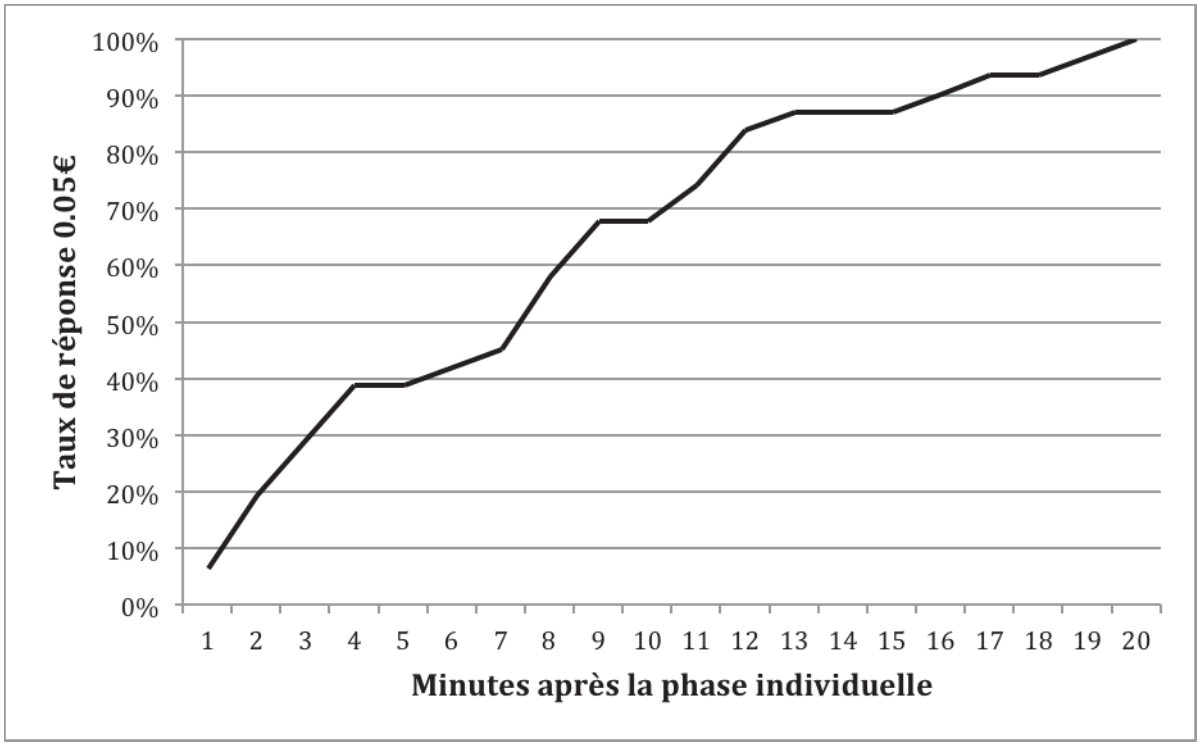
Chapitre 6 - Raisonner en contextes argumentatifs

6.1 La puissance sous-estimée du raisonnement en groupe

Un moyen relativement facile à mettre en place et très efficace pour augmenter les performances des sujets dans des tâches comme celles de Wason est, tout simplement, de faire raisonner les sujets en groupe. Moshman et Geil en 1998 ont montrés que, si seuls 20% des sujets réussissent la tâche individuellement, une fois mis en groupes de 5 ou 6, 80% des sujets obtiennent la bonne réponse.

Dans une des expériences réalisées avant la présente thèse, en collaboration avec Emmanuel Sander et Hugo Mercier, nous avons obtenus des résultats du même ordre pour la tâche de Paul et Linda avec 84 étudiants suisses. Ils passèrent de 10% de réussite en individuel à 63% en groupes de 4 ou 5. De plus la majorité des sujets qui ont adopté la bonne réponse, lors des discussions de groupe, fut capable de réussir les deux tâches de transfert, suggérant qu'ils n'ont pas simplement accepté la conclusion par pression sociale mais ont bien compris les raisons pour leur réponse (Trouche, Sander et Mercier 2014).

Revenons sur les 31 professionnels de la santé dont 29 d'entre eux affichaient une extrême confiance en leur réponse incorrecte au problème du bonbon et de la baguette après 4 minutes de raisonnement individuel. Nous avons cette fois utilisé un paradigme de résolution de problèmes en groupes un peu différent des précédents. A la fin de la phase individuelle, les sujets avaient pour consigne de « se mettre d'accord avec tous leur voisins directs », un sujet pouvant avoir au maximum 8 voisins. Toutes les minutes, je leur demandais d'arrêter leur discussion et de noter leur réponse. Voici, ci-dessous, l'évolution du taux de bonnes réponses à partir de la fin de la phase individuelle.



Au bout de 20 minutes, la bonne réponse s'était diffusée dans toute la salle. Les sujets devaient ensuite justifier leur réponse finale individuellement. Seul deux sujets n'ont pas fourni de justification acceptable.

Malgré l'augmentation remarquable des performances lorsque les sujets résolvent ces tâches en groupe, les psychologues du raisonnement tendent à ignorer, ou en tout cas à sous-estimer ces résultats. Comme nous allons le montrer dans l'article suivant, non seulement la plupart des gens sous-estiment la puissance du raisonnement en groupe, mais c'est le cas aussi de professionnels comme les managers, à priori habitués à gérer des groupes. Pire encore, 32 psychologues considérant le raisonnement comme leur principal domaine d'expertise sont également loin d'estimer les effets de l'argumentation à leur juste valeur. Chaque population fut interrogée sur son estimation du taux de réussite moyen en individuel et en groupe pour la tâche de Wason. Nous avons ensuite calculé leur ratio : taux de réussite estimé en groupe sur taux de réussite estimé en individuel.

Pour la plupart des gens, qu'ils soient simple résidents américains, indiens ou managers européens, le ratio ne dépasse pas 1.5, même après avoir pris connaissance de la bonne réponse à la tâche de Wason. Les 32 psychologues du raisonnement interrogés

font un peu mieux avec un ratio qu'ils estiment à 2.6, mais cela reste tout de même très loin des observations expérimentales, qui indiquent un ratio d'environ 5.

L'article présentant cette étude est issue d'une collaboration avec Hugo Mercier, Hiroshi Yama, Christophe Heintz et Vittorio Girotto et a fait l'objet d'une publication dans la revue *Thinking & Reasoning* en 2014.

Experts and laymen grossly underestimate the benefits of argumentation for reasoning

Hugo Mercier

Université de Neuchâtel

Emmanuel Trouche

CNRS &

Université Lyon 1

Hiroshi Yama

Osaka City University

Christophe Heintz,

Central European University

Vittorio Girotto,

University IUAV of Venice

Accepted in Thinking & Reasoning.

Not proofread – please do not quote.

Abstract

Many fields of study have shown that group discussion generally improves reasoning performance for a wide range of tasks. This article shows that most of the population, including specialists, does not expect group discussion to be as beneficial as it is. Six studies asked participants to solve a standard reasoning problem—the Wason selection task—and to estimate the performance of individuals working alone and in groups. We tested samples of U.S., Indian, and Japanese participants, European managers, and psychologists of reasoning. Every sample underestimated the improvement yielded by group discussion. They did so even after they had been explained the correct answer, or after they had had to solve the problem in groups. These mistaken intuitions could prevent individuals from making the best of institutions that rely on group discussion, from collaborative learning and work teams to deliberative assemblies.

Keywords: Reasoning; group problem solving; argumentation; intuitions about argumentation.

Descartes forcefully put forward a view of reasoning as chiefly aimed at improving individual cognition: "the kind of logic which teaches us to direct our reason with a view to discovering the truths of which we are ignorant." By contrast, argumentation—"a dialectic which teaches ways of expounding to others what one already knows"—only "corrupts good sense rather than increasing it" (Descartes, 1985, p. 186). Nineteenth century scholars of crowd psychology attacked even more fiercely institutions relying on deliberation such as juries and parliaments (e.g. Le Bon, 1897), and their views exerted a considerable influence on many 20th century intellectuals (see Barrows, 1981; Moscovici, 1985).

Other, generally less influential thinkers have suggested that reasoning chiefly serves social functions, notably argumentation, and that deliberation is an effective mean to gain better beliefs (Cattaneo, 1864, see Billig, 1996; Landemore, 2012). Many studies have vindicated this minority view by demonstrating that group discussion often improves reasoning performance. This improvement has been observed in a wide range of tasks in the laboratory—deductive problems (Laughlin & Ellis, 1986; Moshman & Geil, 1998; Trouche, Sander, & Mercier, in press), inductive problems (Laughlin, Bonner, & Miner, 2002), numerical estimations (Minson, Liberman, & Ross, 2011; Sniezek & Henry, 1989), and various work related problems (Blinder & Morgan, 2005; Lombardelli, Proudman, & Talbot, 2005; Michaelsen, Watson, & Black, 1989)—as well as in various other contexts—such as work teams (Guzzo & Dickson, 1996), political discussions (Fishkin, 2009; Mercier & Landemore, 2012), scientific discussions (Dunbar, 1995; Mercier & Heintz, forthcoming; Okada & Simon, 1997), and forecasting groups teams (Mellers et al., 2014; Rowe & Wright, 1996). Group discussion yields similar improvements in different cultures (Mercier, 2011a; Mercier, Deguchi, Van der Henst, & Yama, submitted) and throughout development, starting with preschool children (Doise & Mugny, 1984; Mercier, 2011b; Perret-Clermont, 1980; Slavin, 1995; Smith et al., 2009). These results are robust provided some minimal conditions

are met, such as allowing everyone to express their true opinions (Janis, 1982), and providing an heterogeneous opinion pool (Sunstein, 2002).

Although these results are robust, and, in some cases, old (Bos, 1937; Joubert, 1932; Shaw, 1932), they are not mentioned in current reasoning handbooks (e.g. Manktelow, 2012), and, as we have observed in informal discussions, often surprise the general public as well as specialists. Although the view that reasoning works better in deliberative than individual settings has been empirically vindicated, it does not seem to have become dominant, even among specialists. This potential ignorance of the benefits of group reasoning could have dire practical consequences, leading for instance individuals to neglect collaborative learning as an educational method, to underuse teams in organizations, or even to scorn institutions that rely on deliberations such as juries.

In this article we evaluate people's intuitions about the efficacy of group discussion using the most investigated reasoning problem: the *selection task*, in which participants have to evaluate the truth status of a conditional statement (Wason, 1966). In the following studies, after tackling the standard, abstract version of the task, participants were asked to estimate how many people would solve it on their own, and how many would solve it after discussing it in small groups. These estimates could then be compared to the data in the literature which suggest that fewer than 15% of participants working on their own provide the correct answer (Manktelow, 2012), while about 70% do so after discussing in groups of 3 to 5 individuals (see Table 1).

The existing data and the estimates could be compared in two ways. First, one could compare the absolute levels of performance, to determine whether participants can correctly estimate how many individuals get the right answer individually and in groups. Second, one can compare the relative levels of performance—for instance, the ratio of group to individual performance—to determine whether participants can correctly estimate the improvement

yielded by group discussion. Here we are interested in whether participants can anticipate that group reasoning outperforms individual reasoning, not in whether they can correctly estimate absolute levels of performance. Thus, we focus on the second type of comparison, namely, the ratios of group to individual performance.

Source	% individuals correct after solitary reasoning	% groups correct after group discussion	% individuals correct after group discussion	Ratio of group to individual performance
(Moshman & Geil, 1998 comparison 1)	9%	70%	N/A	7.47
(Moshman & Geil, 1998 comparison 2)	21%	80%	79%	3.75
(Maciejovsky & Budescu, 2007)	9%	50%	N/A	5.71
(Mercier et al., submitted)	20%	65%	64%	3.13
Weighted averages	15%	63%	N/A	4.14

Table 1. Comparison of individual and group performance on the selection task. The ratios were computed using the "% individuals correct in groups" when possible.

Study 1

Method

Participants

25 participants (56% women, M_{Age} = 38.28, *SD* = 11.12) were recruited through the Amazon Mechanical Turk website. Their I.P. addresses indicated that they were in the U.S. In Studies 1 to 3, participants were paid the normal rate for this type of task.

Design

The order of the questions 'estimation of individual performance' and 'estimation of

group performance' was counterbalanced.

Procedure

Participants were given the standard, abstract version of the selection task to tackle. Once they had answered, they were asked to estimate individual performance ("Out of 100 people trying to solve this problem on their own, how many people do you think would give the correct answer?") and group performance ("Out of 100 people trying to solve this problem by discussing in small groups, how many people do you think would give the correct answer?"). As a debiasing procedure, participants were then provided with the correct answer to the selection task and its explanation, and they had to estimate individual and group performance again. Finally, they answered standard demographic questions.

Results and discussion

The order of the individual and group estimation questions did not significantly affect the answers in this study or any of the other studies in which it was counterbalanced (Studies 1 to 5). Hence, this manipulation will not be reported in the other studies.

To compare estimated performance with actual performance, we used the four comparisons of individual to group performance that we could locate in the literature (see Table 1), treating each as an individual data point. This N of 4 renders the statistical tests very conservative. In Study 1, individual performance was estimated to be 65% correct (*SD* = 19.76), significantly higher than actual individual performance ($t(13.8) = 9.72$, $p < .001$).[1] Group performance was estimated to be 72% correct (SD = 23.15), not significantly different from actual performance ($t(7.2) = 0.90$, $p = .40$), but significantly higher than estimated individual performance ($t(24) = -3.15$, $p = .004$). The ratios of estimated group to individual performance ($M = 1.12$, $SD = 0.23$) was significantly lower than the observed ratios ($t(3.0) = -3.95$, $p = .029$). Answers following the debiasing procedure will be discussed below, after

[1] The fractional degrees of freedom stem from the use of t-tests on samples with unequal variance.

Study 5. Table 2 presents the main results and Figure 1 presents the ratios of individual to group performance from the present studies.

	Estimated individual performance	Estimated group performance
S1 (U.S.) Before Feedback	65	72
S2 (U.S.) Before Feedback	66	56
S3 (India) Before Feedback	57	62
S4 (Japan) Before Group	59	76
S5 (Managers) Before Feedback	57	71
S1 (U.S.) After Feedback	39	51
S2 (U.S.) After Feedback	52	49
S3 (India) After Feedback	47	46
S4 (Japan) After Feedback	63	75
S5 (Managers) After Feedback	36	57
S6 (Psychologists)	16	38
Global average	49	59

Table 2. Estimated individual and group performance from all studies.

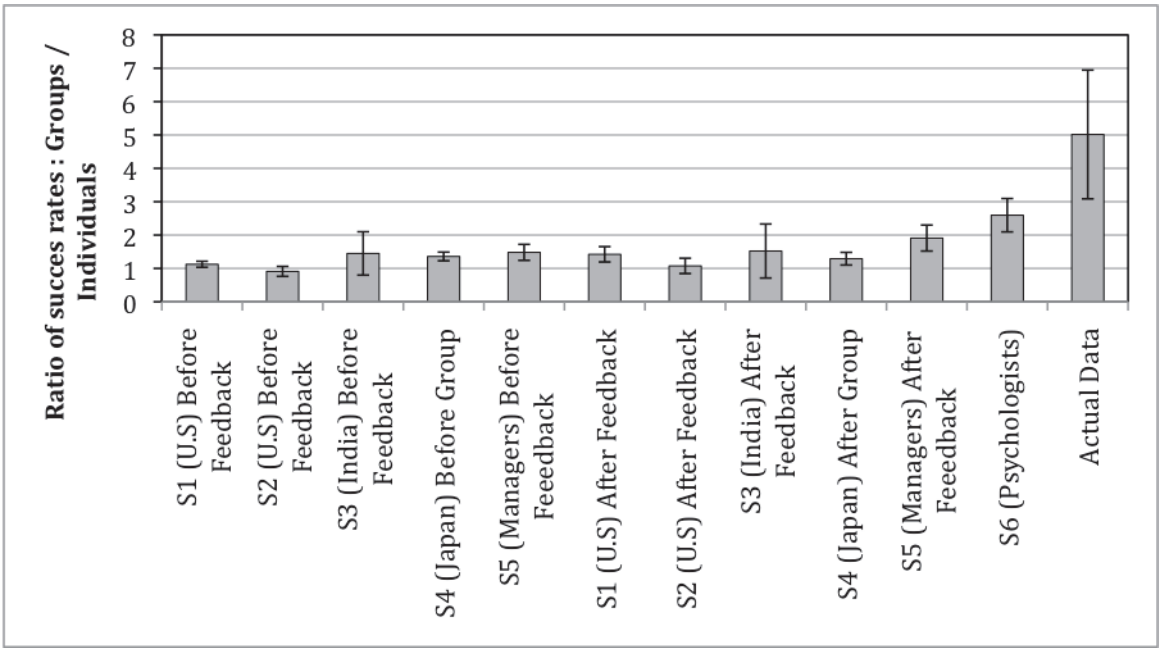


Figure 1. Ratio of group to individual performance: estimates from 6 studies compared with actual data (with 95% confidence intervals).

Previous research has shown that participants fail to appreciate the benefits of aggregating several opinions—averaging opinions in particular—by contrast with choosing one of the opinions (Larrick & Soll, 2006; Soll & Larrick, 2009). In the present case, participants' failure to appreciate the margin by which groups perform better than individuals could stem from difficulties with probabilistic reasoning, namely a failure to compute how many groups would contain at least one member able to find the correct answer on her own. Alternatively, participants could think that even if someone has the correct answer, she will not be able to convince someone with the wrong answer. Given that participants gave a very high estimate of individual performance, the first alternative is unlikely to explain the results, as it would require that there be no mixing of members with the correct and the incorrect answer in any group. Therefore Study 2 focuses on the second explanation, as well as serving as a replication of Study 1.

Study 2

Method

Participants

43 participants (33% women; M_{Age} = 28.0, SD = 4.91) were recruited through the Amazon Mechanical Turk website. They had to be located in the U.S.

Procedure

The procedure was identical to that of Study 1 except that after the estimation questions, participants were asked to directly estimate the effectiveness of argumentation ("Now imagine only two people. One of them has found the correct solution on his or her own, and the other hasn't. The two of them have to agree about an answer. What do you think are the chances that the participant who got the problem right will convince the other? Give an estimate between 0 and 100.").

Results and discussion

The results of Study 1 were replicated. Individual performance was estimated to be 66% correct (SD = 22.18), significantly higher than actual individual performance ($t(11.4) = 10.63$, $p < .001$). Group performance was estimated to be 56% correct (SD = 33.34), not significantly different from actual performance ($t(8.3) = -1.20$, $p = .26$) but significantly lower than estimated individual performance ($t(42) = 2.05$, $p < .05$). As a result, the ratios of estimated group to individual performance ($M = 0.91$, $SD = 0.51$) was significantly lower than the observed ratios ($t(3.0) = -4.15$, $p = .025$).

Participants estimated that someone with the correct answer would convince someone with the wrong answer in 43% ($SD = 25.3$) of the cases. In reality this number is close 100% since the "truth wins" scheme best explains the performance of groups on intellective tasks such as the Wason selection task: As soon as one group member has the correct answer, she nearly always manages to convince the group, even if she is alone and faces a unanimous majority supporting the wrong answer (see for instance Trouche et al., in press). This result confirms that participants grossly underestimate the benefits of argumentation. However, there was no correlation ($r = 0.00$) between the ratios of individual to group performance and the estimation of the efficacy of argumentation in pairs, so that the latter result might not explain the former. Answers following the debiasing procedure will be discussed below.

The underestimation of the benefits of argumentation observed among U.S. participants might reflect the influence of culture specific factors. In particular, Westerners tend to have a more essentialist view of intelligence than Easterners such as than Indian (Rattan, Savani, Naidu, & Dweck, 2012) and Japanese individuals (Heine et al., 2001). An essentialist view of intelligence suggests that intelligence is little affected by learning and other contextual factors (Dweck, 1999) and could therefore explain why U.S. participants do not provide different estimations for individual and group reasoning: they might believe that any individual's chance of providing the correct answer is unaffected by her social setting,

including whether someone else in the group found the correct answer. Accordingly, we replicated Study 1 with participants in India (Study 3) and in Japan (Study 4).

Study 3

Method

Participants

25 participants (36% women; Mage = 33.12, SD = 9.69) were recruited through the Amazon Mechanical Turk website. They had to be located in India.

Procedure

The procedure was identical to that of Study 1 (in English).

Results and discussion

The results of Study 1 were replicated. Individual performance was estimated to be 57% correct (*SD* = 31.17), significantly higher than actual individual performance ($t(23.9) = 5.99$, $p < .001$). Group performance was estimated to be 62% correct (*SD* = 32.07), not significantly different from actual performance ($t(11.6) = -0.44$, $p = .67$) or estimated individual performance ($t(24) = -0.91$, $p = .37$). As a result, the ratios of estimated group to individual performance (*M* = 1.45, *SD* = 1.67) was significantly lower than the observed ratios ($t(3.72) = -3.43$, $p = .030$), and not significantly different from that of U.S. participants ($t(25.25) = 1.36$, $p = .18$). Answers following the debiasing procedure will be discussed below.

Study 4

Method

Participants

35 participants (80% women; *Mage* = 22.8, *SD* = 10.5) took part in the experiment during a class held at Osaka City University. All participants were Japanese. The experiment was not part of the coursework, and students had the option to opt out. The questionnaires

were filled anonymously.

Procedure

The first part of the experiment was identical to that of Study 1, except that the questionnaires were translated into Japanese. Participants had to solve the selection task on their own, and then answer the two estimation questions. As a debiasing procedure, participants were put in small groups and asked to solve the task again. Next, they were asked to answer the two estimation questions again.

Results and discussion

The results of Study 1 were replicated. Individual performance was estimated to be 59% correct (SD = 16.37), significantly higher than actual individual performance ($t(8.1) = 10.19$, $p < .001$). Group performance was estimated to be 76% correct (SD = 17.41), not significantly different from actual performance ($t(4.5) = 1.58$, $p = .18$) but significantly higher than estimated individual performance ($t(34) = -6.77$, $p < .001$). The ratios of estimated group to individual performance (M = 1.36, SD = 0.41) was significantly lower than the observed ratios ($t(3.0) = -3.70$, $p = .034$). However, the ratios were significantly higher than those of U.S. participants (Studies 1 and 2) ($t(73.97) = 4.25$, $p < .001$). Answers following the debiasing procedure will be discussed below.

The results of Studies 1 to 4 suggest that the underestimation of the benefits of argumentation cannot be entirely explained by one critical cultural factor—essentialist thinking about intelligence. They thus suggest that universal mechanisms are at play. However, even if such a cultural factor has no effect on the present results, individual experience with group decision-making might affect the evaluation of individual vs. group reasoning. Managers tend to have extensive experience with team work, and therefore offered a relevant control.

Study 5

Method

Participants

Eighty-six participants took part in the experiment during two classes held at the Central European University (Budapest) as part of an MBA course and an EMBA course. The only data analyzed were from the 46 (35% women; M_{Age} = 35.02, *SD* = 4.51) participants who answered that their current occupation was manager were kept. They had an average of 6.24 years of experience as managers (*SD* = 3.63).

Procedure

Participants answered the questions in a classroom using the online survey of Study 2 with adapted demographic questions (in English).

Results and discussion

The results of Study 2 were replicated. Individual performance was estimated to be 57% correct (SD = 25.29), significantly higher than actual individual performance ($t(13.6) = 8.30$, $p < .001$). Group performance was estimated to be 71% correct (*SD* = 27.25), not significantly different from actual performance ($t(6.10) = 0.80$, $p = .45$), but significantly higher than estimated individual performance ($t(45) = -3.65$, $p < .001$). The ratios of estimated group to individual performance (*M* = 1.48, *SD* = 0.83) were significantly lower than the observed ratios ($t(3.1) = -3.56$, $p = .036$). However, the ratios were higher than those of the non-managers (Studies 1 to 4) ($t(80.88) = -2.10$, $p = .039$) The managers also underestimated the effectiveness of argumentation, judging that someone with the correct answer only had 28 chances out of 100 (*SD* = 20.1) to convince someone with the wrong answer.

Effects of debiasing procedures

The first debiasing procedure, used in Studies 1, 2, 3, and 5, was to explain the correct answer to the selection task. This procedure lowered the estimates of individual (Pre: *M* =

61%, *SD* = 24.8; Post: *M* = 44%, *SD* = 25.8; paired t-test: $t(138) = 7.34$, $p < .001$) and group performance (Pre: *M* = 65%, *SD* = 30.05; Post: *M* = 52%, *SD* = 31.29; paired t-test: $t(138) = 5.47$, $p < .001$). It had little effect on the difference between the estimates of individual to group performance, as this difference remained significant and in the correct direction only for the two groups in which it had the same properties before the debriefing (Study 1, $t(24) = -4.67$, $p < .001$; Study 2, $t(42) = 0.68$, $p = .5$; Study 3, $t(24) = 0.26$, $p = .8$; Study 5 $t(45) = -5.19$, $p < .001$).

The first debiasing procedure had a small but significant positive effect on the ratios of individual to group performance (Pre: *M* = 1.23, *SD* = 0.93; Post: *M* = 1.49, *SD* = 1.30 ; paired t-test: $t(138) = -2.22$, $p = .028$). This effect, however, is entirely driven by the managers (Study 5): Studies 1 to 4 ($t(92) = -1.17$, $p = .244$); Study 5 ($t(45) = -2.37$, $p = .022$). However, the post-debiasing procedure ratios were still significantly lower than the actual ratios ($t(3.08) = -3.55$, $p = .037$), even for the managers ($t(3,2) = -3.09$, $p = .048$).

Explaining to the participants the correct answer had a larger impact on the estimations of the effectiveness of argumentation, possibly because the participants have just been convinced to change their mind in order to adopt the correct answer: Study 2, M_{Pre} = 42%, *SD* = 25.9; M_{Post} = 61%, SD = 21.7, $t(42) = -17.67$, $p < .001$; Study 5, M_{Pre} = 28%, *SD* = 20.3; M_{Post} = 73%, SD = 25.4; $t(45) = -26.76$, $p < .001$. Still, even the post-debiasing estimates were lower than the actual results (close to 100%).

The second debiasing procedure, used in Study 4, was to let participants solve the task in groups. It had no significant effect on the estimates of individual (*M* = 63.29, *SD* = 16.93; paired t-test: $t(34) = -1.28$, $p = .21$), or group performance (*M* = 75.43, *SD* = 15.31; paired t-test: $t(34) = 0.31$, $p = .76$). After the debiasing procedure, the participants still estimated group performance to be higher than individual performance ($t(34) = -3.53$, $p = .001$), but there was no effect of the procedure on the ratios of group to individual performance (*M* =

1.29, *SD* = 0.56; paired t-test: $t(34) = 0.72$, $p = .48$).

The results suggest that the underestimation of the benefits of argumentation is very robust. To further check this conclusion, we tested whether extensive expertise in the psychology of reasoning would allow participants to properly estimate the benefits of argumentation.

Study 6

Method

Participants

Fifty participants were recruited through a professional mailing list (8), personal contacts (27), and at a reasoning workshop (17). We only kept those participants whose self-defined primary field of expertise was psychology of reasoning ($N = 32$) ($M_{Age} = 44.6$, *SD* = *13.9*).

Procedure

Participants were told that the object of the study was the Wason selection task, more specifically the standard, abstract version of the task used in Studies 1 to 5. They were then asked to estimate individual and group performance. Participants then had to estimate the effectiveness of argumentation in a simple debating pair, as in Study 2. Finally they answered some demographic questions.

Results and discussion

Participants correctly estimated individual performance ($M = 16\%$, SD = 10.23; $t(4.96) = 0.29$, $p = .78$), and, while they estimated group performance to be higher than individual performance ($t(31) = -7.28$, $p < .001$)., they still underestimated it ($M = 36\%$, SD = 18.28; $t(4.89) = -4.33$, $p = .008$). As a result, they tended to underestimate the ratio of group to individual performance ($M = 2.60$, SD = 1.43; $t(3.40) = -2.38$, $p = .087$). However, they did so less than the other populations, even after they had been given the correct answer

(comparison with the post-feedback ratios of Studies 1, 2, 3, and 5: $t(43.5) = 3.98$, $p < 0.001$).

The psychologists underestimated the effectiveness of argumentation to the same extent that the participants of Studies 2 and 5, answering that someone with the correct answer would convince someone with the wrong answer in only 68% of the cases (SD = 24.15; $M_{\text{Studies 2 and 5}} = 67\%$, $SD = 24.25$; $t(55.01) = -0.17$, $p = .87$).

This result yields two conclusions. First, even experts in the field who are well acquainted with the individual performance on the selection task do not know of the results demonstrating a dramatic improvement after group discussion. Second, these experts do not have the intuition that such a dramatic improvement would take place.

Conclusion

Participants had to solve a standard reasoning problems (except in one study in which it was already known to the participants), and estimate individual and group performance on the same problem. These estimations were compared to the observed performance of individuals and groups in four experiments. All the groups tested underestimated the increase in performance that follows from group discussion (Figure 1). The ratios of group to individual performance were often close to 1, indicating that on average participants thought group discussion would provide no benefits at all over individual reasoning. Indeed, if we exclude the psychologists, we find that before the debiasing procedure over a third of the participants estimated the performance of groups to be the same or lower than that of individuals (65 out of 177 participants). We obtained convergent results when we asked participants to estimate the effectiveness of argumentation more directly by indicating the chances that someone with the correct answer would convince someone with the wrong answer (Studies 2, 5, and 6).

Besides showing that individuals tend to underestimate the benefits of group discussion, our results also suggest that they overestimate individual performance in this type

of task. The participants even kept overestimating individual performance after they had been explained the correct answer—and thus, for most of them, after realizing that they had given the wrong answer. This phenomenon deserves further investigation.

The first moderator to be studied was culture (Studies 1, 2, 3, and 4). We found that the members of cultures that are supposed to have a less essentialist view of intelligence (Indian and Japanese participants) also grossly underestimated the benefits of argumentation. The Japanese participants did so less than the American participants, but this effect could also depend on other differences between the populations (respectively, students vs. MTurkers) and the experimental settings (respectively, in a classroom vs. online).

The second moderator studied was occupation. In Study 5, the participants were managers, people who have experience working in teams and organizing teamwork. They, too, underestimated the benefits of argumentation, although they did so less than other participants. Again, other factors (such as experimental setting) cannot be entirely ruled out as an explanation for this difference.

The third moderator studied was knowledge of the correct answer, which was manipulated as a within-participant variable. The participants for whom this manipulation had the most effect were the managers, and the question for which this manipulation had the strongest effect was the direct estimation of the chances that someone with the correct answer would convince someone with the wrong answer. The latter result can presumably be explained by the fact that the participants had just been convinced to accept the correct answer themselves, and could therefore more easily imagine how correct arguments can modify beliefs. However, the ratios of group to individual performance were less affected, suggesting that participants failed to translate this understanding of the effectiveness of one to one argumentation into more accurate estimations of group performance.

The fourth moderator, also manipulated as a within-participant variable, was solving

the problem in groups (Study 4). Even though performance significantly improved after group discussion (from 20% to 65% correct), the participants did not provide more accurate ratios of group to individual performance after group discussion. The discrepancy with the effects of the previous moderator might stem from the different sources providing the right answer: the experimenter (who is nearly always believed) vs. other group members (who might convince with less certainty).

Finally, the fifth moderator studied was expertise with the task in hand. In Study 6, participants were psychologists of reasoning, whose knowledge of the task was apparent in their correct estimates of individual performance. However, they grossly underestimated group performance, as well as the chances that someone with the correct answer would convince someone with the wrong answer.

These results demonstrate a consistent underestimation of the benefits of group reasoning. It should be stressed, however, that some participants did indicate that groups would perform better than individuals. In particular, both the managers after they had been given the correct answer, and psychologists of reasoning generated ratios of individual to group performance above 1.5. It is therefore possible that experience with the task in hand, coupled with more general expertise about group reasoning, can lead people to correctly estimate that groups perform better than individuals—while still underestimating the size of this effect, as well as, in the case of the psychologists, the efficacy of argumentation in pairs.

A potential concern with the present study is lack of ecological validity, as one might argue that the Wason Selection Task is not representative of everyday reasoning. The Wason Selection Task was chosen thanks to the robustness of its results both in individuals and in groups, making for a sound benchmark. As noted in the introduction, the benefits of group reasoning extend far beyond this and other demonstrative tasks. It would therefore be worthwhile to conduct similar experiments asking participants to estimate individual and

group performance on other reasoning tasks.

The causes of the underestimation of the benefits of group reasoning should be the topic of further study. In any case, these findings suggest that people might be neglecting argumentation as an effective mean of improving a variety of outcomes, from work decisions to school achievement or even political opinions. None of the investigated moderators enabled participants to provide accurate assessments of the benefits of argumentation. Therefore, our results suggest that explicit teaching on this topic might be necessary in order to counteract people's misleading intuitions. Such education could enable individuals to enjoy more of the benefits of argumentation through collaborative learning, work teams, deliberative assemblies, and other institutions that rely on argumentation.

Finally, we would like to stress that these results ought to be of particular interest to specialists of reasoning. These scholars have deployed a substantial amount of ingenuity and energy in trying to improve reasoning performance. Yet they have paid scant attention to group reasoning—arguably the most efficient way of improving reasoning performance. This neglect has been accompanied by a more general neglect of the social uses of reasoning, in particular argumentation. We hope that by pointing out the robustness of the benefits of group reasoning, and by showing that these benefits are far from being intuitive, we might get scholars to pay more attention to the study of group reasoning and argumentation.

Data availability statement

All the data is available at this URL:
https://sites.google.com/site/hugomercier/online%20data%20Experts%20and%20laymen%20grossly%20underestimate%20the%20benefits%20of%20argumentation%20for%20reasoning.xlsx?attredirects=0

Acknowledgements.

We benefitted from an Ambizione grant from the Swiss National Fund (to H.M.) and a PhD grant from the D.G.A. to E.T and a grant from the Italian Ministry of Research (PRIN2010-RP5RNM) to V.G. We thank Mike Oaksford and two anonymous referees for their helpful comments. We also thank the colleagues who, for once, have played the role of participants in Study 6.

References

Barrows, S. (1981). *Distorting mirrors: Visions of the crowd in late nineteenth-century France*. New Haven: Yale University Press.

Billig, M. (1996). *Arguing and Thinking: A Rhetorical Approach to Social Psychology*. Cambridge: Cambridge University Press.

Blinder, A. S., & Morgan, J. (2005). Are two heads better than one? Monetary policy by committee. *Journal of Money, Credit and Banking*, 37, 789-812.

Bos, M. C. (1937). Experimental study of productive collaboration. *Acta Psychologica*, *3*, 315–426.

Cattaneo, C. (1864). Dell'antitesi come metodo di psicologia sociale. *Il Politecnico*, *20*, 262–270.

Descartes, R. (1985). *The Philosophical Writings of Descartes, vol. 1*. Cambridge: Cambridge University Press.

Doise, W., & Mugny, G. (1984). *The Social Development of the Intellect*. Oxford: Pergamon Press.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & Davidson, J.E. (Eds.), *The nature of insight* (pp. 365–395)). Cambridge: MIT Press.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.

Fishkin, J. S. (2009). *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford: Oxford University Press.

Guzzo, R. A., & Dickson, M. W. (1996). Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology*, *47*(1), 307–338.

Heine, S. J., Kitayama, S., Lehman, D. R., Takata, T., Ide, E., Leung, C., & Matsumoto, H. (2001). Divergent consequences of success and failure in Japan and North America: An investigation of self-improving motivations and malleable selves. *Journal of Personality and Social Psychology*, *81*(4), 599.

Janis, I. L. (1982). *Groupthink* (2nd Rev.). Boston: Houghton Mifflin.

Joubert, G. J. (1932). *Indiwiduele en Kollektiewe Prestasie, 'n dijdrae tot die experimentele groepsigologie.* Amsterdam: Swets en Zeitlinger.

Landemore, H. (2012). *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton: Princeton University Press.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions : Misappreciation of the averaging principle. *Management Science*, *52*, 111–127.

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.

Le Bon, G. (1897). *The crowd: A study of the popular mind*. London: Macmillian.

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.

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.

Manktelow, K. (2012). *Thinking and Reasoning: An Introduction to the Psychology of Reason, Judgment and Decision Making*. Hove: Psychology Press.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., … others. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115.

Mercier, H. (2011a). On the universality of argumentative reasoning. *Journal of Cognition and Culture*, *11*, 85–113.

Mercier, H. (2011b). Reasoning serves argumentation in children. *Cognitive Development*, *26*(3), 177–191.

Mercier, H., Deguchi, M., Van der Henst, J.-B., & Yama, H. (submitted). The benefits of argumentation are cross-culturally robust: The case of Japan.

Mercier, H., & Heintz, C. (forthcoming). Scientists' argumentative reasoning. *Topoi*.

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*, *33*(2), 243–258.

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.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to Tango. *Personality and Social Psychology Bulletin*, *37*(10), 1325–1338.

Moscovici, S. (1985). *The age of the crowd: A historical treatise on mass psychology*. Cambridge: Cambridge University Press.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning*, *4*(3), 231–248.

Okada, T., & Simon, H. A. (1997). Collaboration discovery in a scientific domain. *Cognitive Science*, *21*(2), 109–146.

Perret-Clermont, A.-N. (1980). *Social Interaction and Cognitive Development in Children*. London: Academic Press.

Rattan, A., Savani, K., Naidu, N. V. R., & Dweck, C. S. (2012). Can everyone become highly intelligent? Cultural differences in and societal consequences of beliefs about the universal potential for intelligence. *Journal of Personality and Social Psychology*, *103*(5), 787.

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, *12*(1), 73–89.

Shaw, M. E. (1932). A comparison of individuals and small groups in the rational solution of complex problems. *The American Journal of Psychology*, *44*(3), 491–504.

Slavin, R. E. (1995). *Cooperative Learning: Theory, Research, and Practice* (Vol. 2nd). London: Allyn and Bacon.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*(5910), 122.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*(1), 1–28.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, *10*(2), 175–195.

Trouche, E., Sander, E., & Mercier, H. (in press). Arguments, more than Confidence, Explain the Good Performance of Reasoning Groups. *Journal of Experimental Psychology: General*.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology: I* (pp. 106–137). Harmandsworth, England: Penguin.

6.2 Expliquer les échecs et les réussites du raisonnement en groupe

Comment expliquer la différence entre les échecs individuels et la réussite des groupes dans les tâches de raisonnement ? Nous avons précédemment expliqué les mauvaises performances du raisonnement individuel par la fonction argumentative du raisonnement. Expliquons brièvement les raisons des réussites en groupe avant d'en exposer les limites.

Lors d'une discussion en groupe les sujets sont tous capables de produire des raisons pour leur point de vue grâce au biais vers son côté . De plus, ces raisons sont ensuite évaluées de façon non-biaisée au sens où, comme le prédit le cadre théorique de la vigilance épistémique, les arguments peuvent surpasser le manque de confiance en la source ou un fort sentiment de confiance chez le récepteur. C'est ce que nous avons testé dans l'étude expérimentale suivante. Nous avons fait résoudre le problème de Paul et Linda et une version modifié du bonbon et de la baguette à des sujets américains recrutés en ligne grâce à la plateforme Amazon Mechanical Turk ainsi que Paul et Linda seulement à des sujets Chinois recrutés grâce à la plateforme Sojump. Ne conservant que les sujets donnant la mauvaise réponse intuitive, nous leur avons donné un argument pour la bonne réponse en leur disant qu'il venait d'un précédent participant. Certains sujets ont reçu des informations concernant ce participant. Pour faire varier la compétence de la source, il était dit que le participant dont ils allaient recevoir un argument avait réussi auparavant 0(zéro) problème de raisonnement sur 8 (Condition source incompétente) ou qu'il avait réussi 8 problèmes de raisonnement sur 8 (Condition source compétente). Nous avons également fait varier la bienveillance de la source, indiquant au participant que le sujet dont ils allaient recevoir l'argument était motivé financièrement pour que les gens qui reçoivent son argument se trompent (Condition source malveillante) ou à l'inverse qu'il était motivé financièrement pour mener les gens à la bonne réponse (Condition source bienveillante). De plus certains sujets n'ont pas reçu d'information concernant la source mais ont dû répondre au problème le plus rapidement possible (Condition faible confiance en soi) ou à l'inverse les sujets étaient encouragés à prendre le temps de réfléchir à leur réponse et il leur était demandé de la justifier (Condition haute confiance en soi). Comme le montrent les

résultats, le taux d'acceptation de l'argument reste remarquablement stable à travers les conditions.

Cette étude est le fruit d'une collaboration avec Hugo Mercier et Shao Jing, elle n'a pas encore fait l'objet d'une publication.

How is argument evaluation biased?

Trouche, Emmanuel

CNRS

Laboratoire Langage, Cerveau et Cognition

67, Boulevard Pinel, 69675 Bron, France

Shao, Jing

EPHE & Université Paris 8

Laboratoire Cognitions Humaine et Artificielle

4-14, rue Ferrus, 75014 Paris, France

& Université de Haute-Alsace

2, rue des Frères Lumière, 68093 Mulhouse, France

Mercier, Hugo

Université de Neuchâtel

Centre de Sciences Cognitives

Espace Louis-Agassiz 1 Neuchâtel 2000

Abstract

Results suggest that participants are more critical of arguments that challenge their views or that come from untrustworthy sources. This bias, however, might only occur because participants are prone to generate many counter-arguments in response to such arguments. The immediate evaluation that takes place as people read or hear an argument might not be biased. In four experiments, we presented participants with an argument for which there is no good counter-arguments. In Experiments 1, 3, and 4, participants were made to be particularly confident in the answer challenged by the argument. This did not make them less likely to accept the argument. In Experiments 2, 3, and 4, participants were made to trust or distrust the source of the argument. Source trustworthiness had relatively limited effects on argument acceptation. Overall, our results support the hypothesis that the observed biases in argument evaluation are restricted to delayed argument evaluation, while immediate argument evaluation might be free of such biases. In conclusion, we spell out some methodological, theoretical, and practical consequences of this hypothesis.

Keywords: Argument evaluation; prior beliefs; trust; bias.

How is argument evaluation biased?

Given the amount of information people deal with daily in modern societies, the ability to properly evaluate arguments has become an even more critical skill. Not only should we be able to reject fallacies, but we should also be able to accept strong arguments, even if they challenge our prior beliefs or if they come from people who we don't trust.

Experimental results suggest that argument evaluation does not behave as expected. People seem to be able to evaluate argument strength appropriately in the absence of other factors (see, e.g., Hahn & Oaksford, 2007; Hoeken, Šorm, & Schellens, 2014; Hornikx & Hahn, 2012). However, when people are not highly motivated, information about the source of the message seems to taint the way they evaluate arguments (Petty & Wegener, 1998). Moreover, individuals seem to evaluate more critically arguments whose conclusion challenge strongly held beliefs (Edwards & Smith, 1996).

Here we will refer to the effects of trust in the source of the arguments and prior beliefs in the argument's conclusion as factors that *bias* argument evaluation. Bias often has a negative connotation in this context, but here we will use it as a neutral way of describing the influence of these factors on argument evaluation.

The goal of this article is threefold. First, to point out a significant difference between two meanings of 'argument evaluation' that, we argue, correspond to the operation of different cognitive mechanisms. We will refer to these two meanings as *immediate argument evaluation* and *delayed argument evaluation*. The second goal is to review the literature in order to show that there is strong evidence that delayed argument evaluation is biased, but not that immediate argument evaluation is. The third goal is to offer novel evidence suggesting that immediate argument evaluation might not be biased.

Immediate versus delayed argument evaluation

When people evaluate arguments, whether it is in the context of a laboratory experiment or in everyday life, several cognitive mechanisms can be at play. Arguably, the first step of argument evaluation takes place as people process and attempt to understand the argument. We will call this step *immediate argument evaluation*. For the vast majority of everyday arguments, which are simple, immediate argument evaluation is very fast, possibly as fast as the processing of any other utterance. That immediate argumentation takes place very quickly is shown by the fact that people can exchange arguments at a fast pace, exchanges that often require an understanding of the arguments one is replying to (see, e.g. Resnick, Salmon, Zeitz, Wathen, & Holowchak, 1993). It is entirely possible that immediate argument evaluation takes place at the subpersonal level (Dennett, 1991), so that people might not have

full access to why they found an argument to be more or less convincing—in the same way as one might be unable to explain why a statement is ungrammatical.

Directly assessing immediate argument evaluation is difficult. Most psychology experiments ask for explicit ratings of arguments and they give participants time to reflect on their answers. As a result, they measure what we will call *delayed argument evaluation*. Many studies have supposedly demonstrated biases in delayed argument evaluation. Many of these studies have relied on indirect measures of argument evaluation. Studies relying in indirect measures do not ask participants to rate arguments directly; instead they ask participants to rate an argument's conclusion (e.g., Evans, Newstead, & Byrne, 1993) or to evaluate the impact of an argument by comparing the participants' attitudes before and after presentation of the argument (e.g., Petty & Wegener, 1998). They do not directly ask participants to evaluate the strength of an argument. These studies have revealed that participants' acceptance of an argument's conclusion is determined not only by argument strength, but also by other factors such as the participants' prior beliefs in the argument's conclusion and their trust in the argument's source. This should not be a surprising outcome. What matters when deciding whether to accept a given conclusion should not only be the strength of the arguments supporting it. The confidence people have in the conclusion, as well as the trustworthiness of its source should also be taken into account (see, e.g., Mercier, submitted; Sperber et al., 2010). These results thus do not directly demonstrate that delayed argument evaluation is biased, they only show that other cues are integrated with argument evaluation to form an overall evaluation of the conclusion.

Direct measures of argument strength offer stronger evidence that delayed argument evaluation is biased. They consist in asking participants to evaluate the strength of arguments. Many experiments have shown factors besides argument strength influence participants' evaluation of argument strength (Edwards & Smith, 1996; Klaczynski & Gordon, 1996; Munro et al., 2002; Taber & Lodge, 2006).

Lord *et al.* (1979) offers a classic demonstration of the influence of prior beliefs on delayed argument evaluation. Participants were asked to evaluate studies purportedly supporting a pro and an anti death penalty conclusion. Even though the studies were identical except for the conclusion reached, the participants evaluated studies that did not support their own position more negatively than studies supporting their position. Hahn *et al.* (2009) provides a demonstration bearing on the influence of trust in the source. Participants were asked to evaluate arguments about the safety of various drugs. The argument ratings were shown to depend not only on argument strength, but also on the reliability of the argument's source.

Immediate versus delayed argument evaluation

The studies reviewed above demonstrate that delayed argument evaluation is biased by participant's prior beliefs about an argument's conclusion and by the estimated trustworthiness of its source. Since immediate argument evaluation takes place before delayed argument evaluation, these biases could result from biases in immediate argument evaluation, biases in delayed argument evaluation, or a combination of both.

A well-established source of bias in delayed argument evaluation is the generation of counter-arguments. For instance, Edwards and Smith (1996) asked participants who had an opinion on the matter (here, supporting or opposing the death penalty) to rate arguments such as the following:

> Sentencing a person to death ensures that he/she will never commit another crime. Therefore, the death penalty should not be abolished. (Edwards & Smith, 1996, p. 9)

Results showed that participants rated less well arguments whose conclusion they opposed ('incompatible' arguments). The results also showed that participants took more time reading incompatible rather than compatible arguments, and that they generated more thoughts about incompatible than compatible arguments. The thoughts generated were biased for compatible and for incompatible arguments: in both cases, participants tended to generate thoughts that were congruent with their point of view. However, because more thoughts were generated in response to incompatible arguments, participants ended up generating many more thoughts critical of incompatible arguments than thoughts supportive of compatible arguments.

As suggested by the longer reading times and the greater quantity of counter-arguments eventually produced, participants could thus have generated counter-arguments after reading incompatible arguments and before rating them. These counter-arguments could have caused a bias in delayed argument evaluation. After all, it is quite sensible to rate less well an argument for which one has found many counter-arguments. Other studies also show a strong link between the generation of counter-arguments and the low ratings given to arguments that challenge our prior beliefs (e.g., Greenwald, 1968; Taber & Lodge, 2006).

If this interpretation is correct, the question becomes: Why do people generate more counter-arguments when they disagree with an argument's conclusion than when they agree with it? The arguments used in these studies should rarely lead an individual to change her mind altogether. For instance, few critics of the death penalty would be completely swayed by the argument above. If an argument doesn't persuade someone, the normal reaction is to produce counter-arguments in order to justify one's rejection of the argument's conclusion. By contrast, accepting a conclusion requires less elaboration.

If this interpretation is correct, individuals could evaluate arguments relatively objectively. They would then aggregate the information about argument strength to other cues—such as prior beliefs in the conclusion and trustworthiness of the source. As a result of this aggregation, conclusions that challenge their views, or that are defended by untrustworthy sources, would be less likely to be accepted. If an individual rejects a conclusion, she will likely generate counter-arguments to justify this rejection. These counter-arguments would then affect the delayed evaluation of the argument. If this is the case, then immediate argument evaluation needs not be biased. We will come back in the conclusion to the methodological, theoretical, and practical significance of the potential absence of bias in immediate argument evaluation.

The present experiments

The goal of the present experiments is to offer more evidence of the role of counter-arguments in argument evaluation. To do so, we used deductive arguments. These arguments are not always understood but, once they are, they do not accept good counter-arguments. If it is true that biases in argument evaluation are mostly biases in delayed argument evaluation resulting from the generation of counter-arguments, then these biases should largely disappear when such arguments are used.

Here is an example of the problems we used, taken from Levesque (1986), and which we call the Paul and Linda problem:

> Paul is looking at Linda and Linda is looking at Patrick. Paul is married but Patrick is not. Is a person who is married looking at a person who is not married?
> Yes / No / Cannot be determined.

For such problems, we can confront participants with arguments for the correct answer for which there is no good counter-argument, such as the following:

> Linda is either married or not married. If she is married, then she is looking at Patrick, who is not married, so the answer is Yes. If she is not married, then Paul, who is married is looking at her, so the answer is Yes again. So the answer is always Yes.

In the present experiments, participants were asked to tackle this or a similar problem; they were then given the argument for the correct answer and asked if they wanted to revise their answer, thereby providing us with

a measure of how they had evaluated the argument. The only participants of interest will be those who provide the intuitive wrong answer, given that those who provide the right answer do not change their mind when confronted with an argument for the wrong answer.

In Experiment 1 we manipulated how confident participants were in their answers. In Experiment 2 we manipulated how trustworthy the source of the argument was. Experiment 3 replicated Experiments 1 and 2 and in different culture. Experiment 4 replicated Experiments 1 and 2 with a different problem.

Experiment 1

In a previous experiment, participants had been given the Paul and Linda problem, asked to rate their confidence, provided with an argument for the correct answer, and offered a chance to change their mind (Trouche, Sander, & Mercier, 2014, Experiment 2a). A median split of confidence indicated that, among the participants providing the intuitive wrong answer (N=166), those who were more confident were not less likely to accept the argument (42%) than those who were more confident (36%).[2] This result suggests that, in this case, being confident that an argument's conclusion is incorrect does not make one less likely to accept the argument.

These results, however, are correlational. Participants who think they are relatively good at solving this type of problem might be very confident, even if in fact they have the wrong answer. It is thus possible that participants' general ability to solve this type of problem positively affects both their confidence and their likelihood of understanding the argument for the correct answer.

In order to provide more evidence that confidence in the wrong answer does not affect the chances that the argument for the correct answer is accepted, in Experiment 1 confidence was experimentally manipulated.

Method

Participants

148 participants were recruited online using Amazon Mechanical Turk (52 females, M_{age} = 29.7, SD = 8.5). They were paid 0.70$ for their participation and had to be located in the U.S.

Materials and procedure

Participants were asked to complete the Paul and Linda problem. They then had to evaluate their confidence in the correctness of their answer by choosing from eight indicators of confidence. To avoid ceiling

[2] Each time we analyze median splits, we have a choice to put the participants whose answer falls on the median in either category. We always report the results that are the less favorable to the hypothesis that immediate argument evaluation is not biased.

effects due to overconfidence, the scale was skewed to include answers denoting strong confidence (inspired by Kuhn & Lao, 1996) so that it ranged from "Not confident at all" to "As confident as in the things I'm most confident about" (see ESM). The participants were then provided with an argument for the correct answer, presented as coming from a previous participant, and they had to tackle the problem again. Finally, the participants had to solve two transfer problems, one structurally identical to the initial problem, and one with a similar form but designed so that the correct answer would be the intuitive but wrong answer to the initial problem (see ESM).[3]

In the *Low Confidence* condition participants were told, before completing the problem, that they had to answer quickly. In the *High Confidence* condition participants were told instead that they had to think very carefully about their answer and that they would have to justify it. They were indeed asked to justify their answer. Previous results suggest that thinking about their answer and having to provide reasons for it is likely to make participants more confident (Koriat, Lichtenstein, & Fischhoff, 1980; Tesser, 1978).

Results and discussion

Manipulation check. All confidence results were turned into percentages. The confidence manipulation successfully increased confidence for all participants (data in ESM), and also for the relevant participants, those who provided the intuitive wrong answer (*Low Confidence*, N = 42, confidence = 64% [*Mdn* = 71, *SD* = 27]; *High Confidence*, N = 45, confidence = 76% [*Mdn* = 86, *SD* = 23]; $W = 693$, $p = .029$).

Effects of the manipulation. In the *Low Confidence* condition, 70% of the relevant participants changed their mind; 80% did so in the *High Confidence* condition (Fisher exact test, $p = .32$, $OR = 0.56$) (Figure 1).

These results confirm that, for this type of argument, being confident that an argument's conclusion is incorrect does not make one less likely to accept the argument. Argument evaluation is thus not biased by how confident participants are in the belief that is challenged by the argument.

[3] Although transfer problems were included in all the present experiments, in order not to burden the result section, the results from the transfer problems will be discussed when they are most relevant, namely in a section following Experiment 4 that discusses two possible interpretation of the results. All the data are available in the ESM.

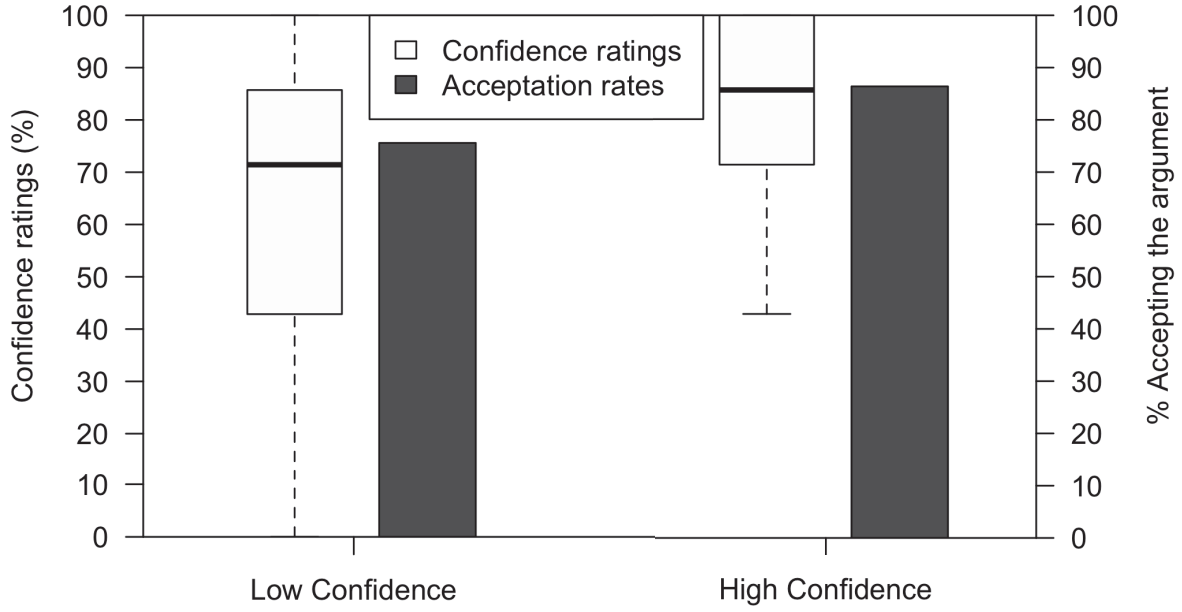


Figure 1. Confidence ratings and percentage of participants accepting the argument for the correct answer in the two conditions of Experiment 1.

Experiment 2

Method

Participants

500 participants were recruited online through Amazon Mechanical Turk (181 females, M_{age} = 31.6, *SD* = 10.2). They were paid 0.70$ for their participation and had to be located in the U.S.

Materials and procedure

Participants were asked to complete the Paul and Linda problem, to provide a justification for their answer, and to evaluate their confidence in the correctness of their answer. They were then told that they would be given the answer and the argument of a participant from a previous experiment. The manipulation bore on how this participant was introduced. In the *Control* condition, the introduction was neutral:

> We are now going to give you the answer that was given by another participant, along with his or her justification.

After this short introduction, which was common to all conditions, the participants could read, in the *High*

Honesty condition:

> When we gathered these results, we were doing a special experiment. In this experiment, the participants were told: YOU SHOULD TRY TO BE AS HELPFUL AS POSSIBLE TO THE OTHER PARTICIPANT AS POSSIBLE. As a result, people were very careful to try and give the right answer and the right justification.

In the *Low Honesty* condition:

> However, you should be careful. When we gathered these results, we were doing a special experiment. In this experiment, the participants were told: YOU SHOULD TRY TO MISLEAD ANOTHER PARTICIPANT WITH YOUR ANSWER AND YOUR JUSTIFICATION. Moreover, they were told that they would be getting a bonus of $1 whenever they managed to make someone adopt the wrong answer.

In the *High [resp. Low] Competence* condition:

> When we gathered these results, we were asking people to answer a series of 8 other problems similar to this one. We can tell you that the participant whose answer you are going to see got
> 8 [resp. 0]
> problems right out of 8 (not including the present problem obviously).

Participants were then asked, "How much do you trust the other person to give you the correct answer," which they could answer on a sliding scale from 0 ('Impossible') to 100 ('Absolutely certain'). Participants in all conditions were then provided with the correct answer and an argument for the correct answer, and they were asked to tackle the problem again. Finally, they had to solve the same transfer problems as in Experiment 1.

Results

Manipulation check. The trust manipulation was successful among all participants (data in ESM) as well as among the participants who provided the intuitive wrong answer. Trust was higher in the *High Honesty* condition (N = 71, trust = 67% [*Mdn* = 70, *SD* = 16]) than in the *Control* condition (N = 70, trust = 58% [*Mdn* = 59, *SD* = 22]; $W = 1898$, $p = .015$), and the *Low Honesty* condition (N = 71, trust = 17% [*Mdn* = 10, *SD* = 23]; $W = 337$, $p < .001$). It

was also higher in the *Control* condition than in the *Low Honesty* condition (*W = 601, p < .001*). Trust was higher in the *High Competence* condition (N = 76, trust = 83% [*Mdn* = 90, *SD* = 17]) than in the *Control* condition (*W = 917, p < .001*), and the *Low Competence* condition (N = 73, trust = 12% [*Mdn* = 1, *SD* = 22]; *W = 140, p < .001*). It was also higher in the *Control* condition than in the *Low Competence* condition (*W = 466, p < .001*).

Effects of the manipulation. Fisher exact tests indicate that participants in the *Control* condition were not significantly more or less likely to accept the argument than those in any other condition (N = 70, 70% accept; *High Honesty* condition, N = 69, 74% accept, *p* = .71, *OR* = .82; *Low Honesty* condition, N = 69, 62% accept, *p* = .37, *OR* = 1.4; *High Competence* condition, N = 74, 81% accept, *p* = .17, *OR* = .55; *Low Competence* condition, N = 72, 61% accept, *p* = .29, *OR* = 1.5).[4]

Participants in the *High Honesty* condition were not significantly more likely to accept the argument than those in the *Low Honesty* condition (*p* = .20, *OR* = 1.7). Participants in the *High Competence* condition were significantly more likely to accept the argument than those in the *Low Competence* condition (*p* = .01, *OR* = 2.71) (Figure 2).

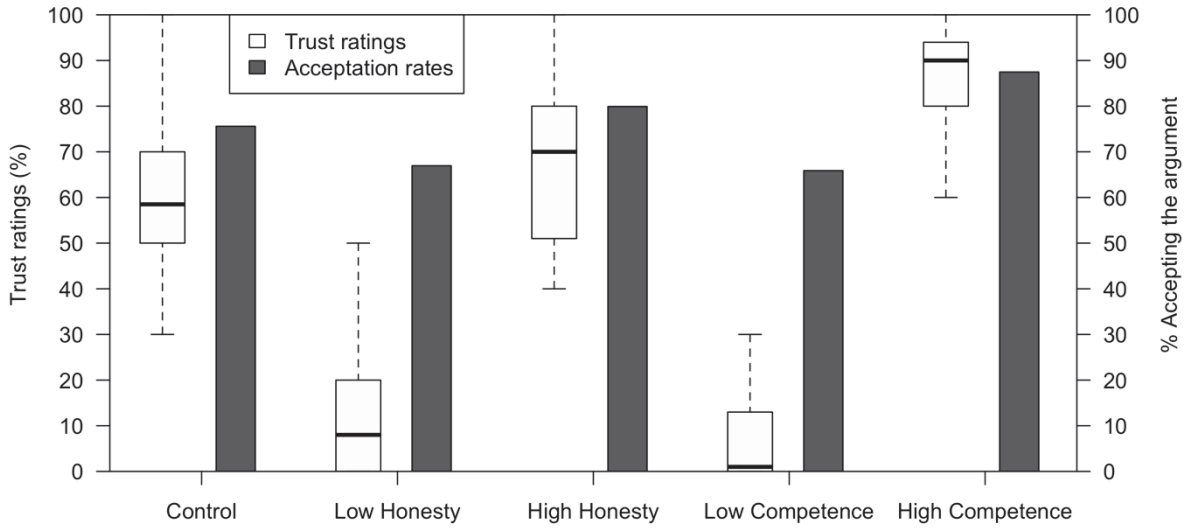


Figure 2. Trust ratings and percentage of participants accepting the argument for the correct answer in the five conditions of Experiment 2.

Discussion

In this experiment, source trustworthiness influenced the likelihood that the argument was accepted:

[4] In this and all subsequent analysis, we excluded the very few participants who neither kept their initial wrong answer nor accepted the correct answer (there were never more than two such participants in each condition) (see ESM for the exact numbers).

arguments from trustworthy sources were more likely to be accepted than argument from untrustworthy sources. However, this effect of source trustworthiness was relatively modest. When asked to evaluate the source's trustworthiness, participants in the control condition gave answers that differed from those of each experimental condition. By contrast, the participants in the control condition did not differ from those of any of the experimental condition in their likelihood of accepting the argument. In particular, even in the low trust conditions, when the participants indicated that they did not trust the source to provide the correct answer, they were still more likely than not to accept its argument (see Figure 2).

Experiments 3: Replication in a different culture

In order to test the robustness of the results above we replicated Experiments 1 and 2. Moreover, in order to establish a degree of cross-cultural robustness, the replication was carried out in China. Easterners, relative to Westerners, have been shown to pay more attention to the context compared to a focal object (Choi, Nisbett, & Norenzayan, 1999; Miyamoto, Nisbett, & Masuda, 2006; Nisbett & Miyamoto, 2005). In the present study, the content of the argument can be seen as the focal object, while one's prior belief in the conclusion and the source trustworthiness can be seen as belonging to the context (Norenzayan, Smith, Kim, & Nisbett, 2002). As a result, Chinese participants could put more weight on prior beliefs and source trustworthiness, and less weight on argument strength, than Westerners.

A pre-test (N = 57) established that the Paul and Linda problem had the required properties in the relevant population: most participants (70%) provided the intuitive but wrong answer; participants were much more likely to accept an argument for the correct answer (65%) than an argument for the wrong answer (18%); the percentage of participants accepting the correct argument avoided floor and ceiling effects (see ESM for details).

Method

Participants

461 participants were recruited online through the crowdsourcing website Sojump (247 females, Mage = 30.5, SD = 6.64). They were paid 1.5 Yuan for their participation and had to be located in the People's Republic of China.

Materials and procedure

Experiment 3 comprises all the conditions of Experiments 1 and 2 except for the *Control Condition* of Experiment 2, for a total of six conditions: *High Confidence*, *Low Confidence*, *High Honesty*, *Low Honesty*, *High Competence*, *Low Competence*. All the materials were translated in Chinese by one of the authors and back translated using the online service Genko. The minor discrepancies that resulted were easily resolved.

Results of the confidence manipulation

Manipulation check. Confidence results were translated into percentages. The confidence manipulation successfully created a difference in confidence for all participants (data in ESM), and also for the relevant participants, those who provided the intuitive wrong answer (*Low Confidence*, N = 37, confidence = 67% [*Mdn = 71, SD = 27*]; *High Confidence*, N = 42, confidence = 82% [*Mdn* = 86, *SD* = 21]; *W = 526, p = .011*) (Figure 3).

Effects of the manipulation. In the *Low Confidence* condition, 61% of the relevant participants changed their mind; 71% did so in the *High Confidence* condition (Fisher exact test, *p* = .35, OR = 0.63) (Figure 3).

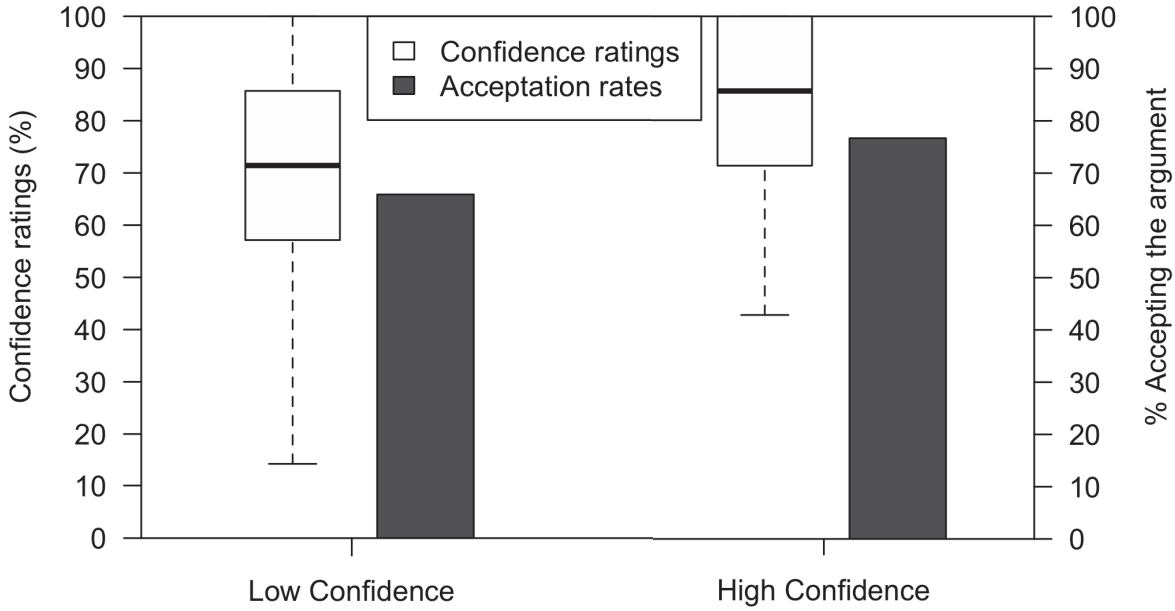


Figure 3. Confidence ratings and percentage of participants accepting the argument for the correct answer in the two conditions of the confidence manipulation of Experiment 3.

Results of the trust manipulation

Manipulation check. The trust manipulation was successful among all participants (data in ESM) as well as among the participants who provided the intuitive wrong answer. Trust was higher in the *High Honesty* condition (N = 51, trust = 73% [*Mdn* = 75, *SD* = 20]) than in the *Low Honesty* condition (N = 55, trust = 41% [*Mdn* = 42, *SD* = 29]), *W = 501, p < .001*). Trust was higher in the *High Competence* condition (N = 47, trust = 74% [*Mdn* = 80, *SD* = 17]) than in the *Low Competence* condition (N = 46, trust = 37% [*Mdn* = 30, *SD* = 30]), *W = 354, p < .001*). Figure 4 displays the trust ratings and the percentage of participants accepting the correct argument in each condition.

Effects of the manipulation. Participants in the *High Honesty* condition (N = 51, 57% accept) were not more likely to accept the argument than those in the *Low Honesty* condition (N = 54, 56% accept, stats). Participants in the *High Competence* condition (N = 46, 70% accept) were not more likely to accept the argument than those in the *Low Competence* condition (N = 45, 67% accept, *p* = .82, *OR* =1.14). Pooled together, the participants in the *High Trust Group* (N = 97, 63% accept) were not more likely to accept the argument than those in the *Low Trust Group* (N = 99, 61% accept, *p* = .77, *OR* =1.10) (Figure 4).

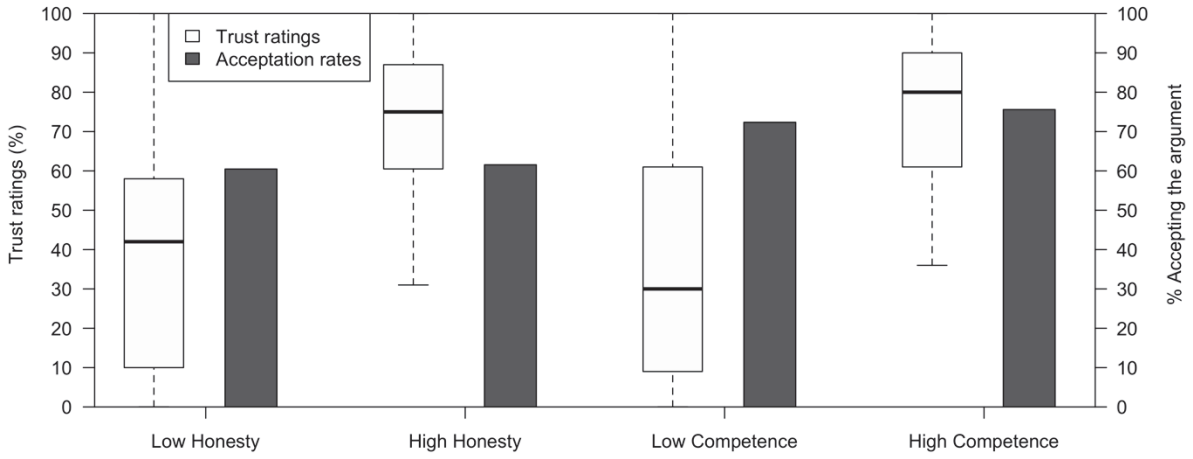


Figure 4. Trust ratings and percentage of participants accepting the argument for the correct answer in the five conditions of the trust manipulation of Experiment 3.

Discussion

The experiment successfully manipulated the confidence participants had in their own answer, and the trustworthiness of the argument's source. These manipulations did not significantly affect argument evaluation, suggesting that argument evaluation was not biased by these factors.

Experiment 4: Replication with a different problem

To ensure that the results of the previous experiments are not restricted to the Paul and Linda problem, we replicated them using a different problem, a variation on the bat and ball problem from the Cognitive Reflection Test (Frederick, 2005):

A trial jury is composed of 12 people. There are 8 more women than men in this jury. How many men are in the jury?

2 / 4 / 6

The argument for the correct answer we used was:

It says that there are 8 more women than men, so if there are 2 men, there are 2 + 8 women, so 10 women. And 2 men plus 10 women makes 12 people jury.

A pre-test (N = 50) established that this problem had the required properties: many participants (52%) provided the intuitive but wrong answer; participants were much more likely to accept an argument for the correct

answer (35%) than an argument for the wrong answer (0%); the percentage of participants accepting the correct argument avoided floor and ceiling effects (see ESM for details).

Method

Participants

452 participants were recruited online through Amazon Mechanical Turk (178 females, *M*age = 33, *SD* = 10.9). They were paid 0.70$ for their participation and had to be located in the U.S.

Materials and procedure

The materials and procedure were exactly identical to those of Experiment 3, except that the problem used was different (and the experiment was conducted in English).

Results confidence manipulation

Manipulation check. All confidence results were translated into percentages. The confidence manipulation successfully created a difference in confidence for all participants (data in ESM), and also for the relevant participants, those who provided the intuitive wrong answer (*Low Confidence*, N = 61, confidence = 66% [*Mdn* = 71, *SD* = 30]; *High Confidence*, N = 42, confidence = 82% [*Mdn* = 100, *SD* = 23]; *W* = 870, *p* = .005).

Effects of the manipulation. In the *Low Confidence* condition, 48% of the relevant participants changed their mind; 43% did so in the *High Confidence* condition (Fisher exact test, *p* = .69, *OR* = 1.2) (see Figure 5).

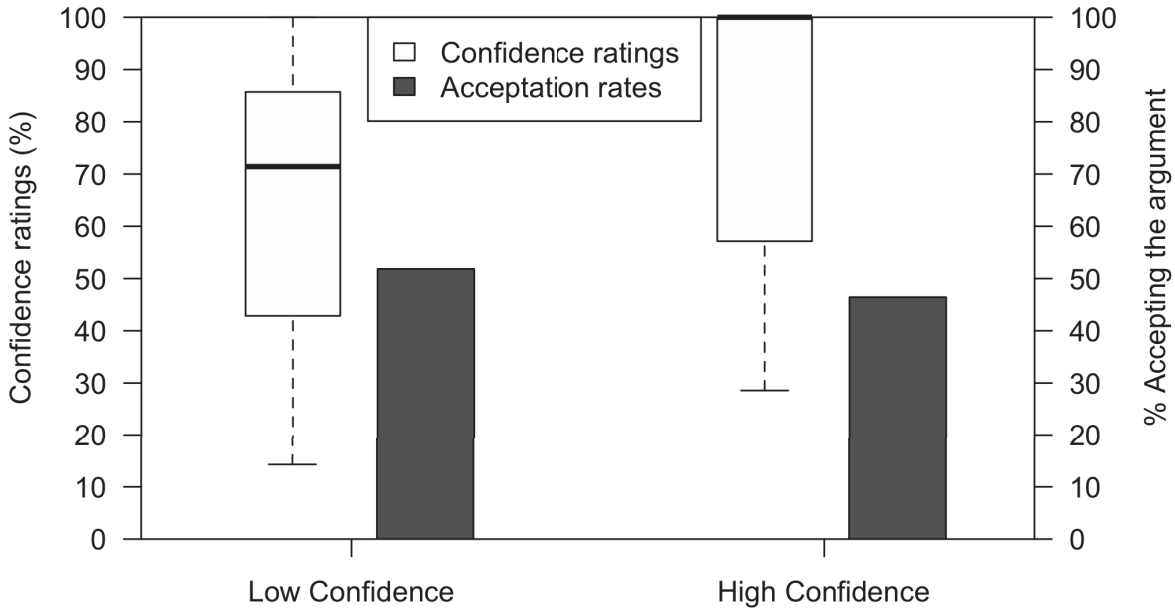


Figure 5. Confidence ratings and percentage of participants accepting the argument for the correct answer in the two conditions of the confidence manipulation of Experiment 4.

Results trust manipulation

Manipulation check. The trust manipulation was successful among all participants (data in ESM) as well as among the participants who provided the intuitive wrong answer. Trust was higher in the *High Honesty* condition (N = 40, trust = 69% [*Mdn* = 71, *SD* = 22]) than in the *Low Honesty* condition (N = 44, trust = 19% [*Mdn* = 1, *SD* = 30]), $W = 205$, $p < .001$). Trust was higher in the *High Competence* condition (N = 35, trust = 67% [*Mdn* = 84, *SD* = 38]) than in the *Low Competence* condition (N = 43, trust = 9% [*Mdn* = 3, *SD* = 17]), $W = 220$, $p < .001$)

Effects of the manipulation. Participants in the *High Honesty* condition (N = 40, 45% accept) were more likely to accept the argument than those in the *Low Honesty* condition (N = 42, 12% accept, $p = .001$, *OR* =5.9). Participants in the *High Competence* condition (N = 35, 54% accept) were not more likely to accept the argument than those in the *Low Competence* condition (N = 43, 42% accept, $p = .36$, *OR* =1.6). Pooled together, the participants in the *High Trust Group* (N = 75, 49% accept) were more likely to accept the argument than those in the *Low Trust Group* (N = 85, 27% accept, $p = .005$, *OR* =2.6) (Figure 6).

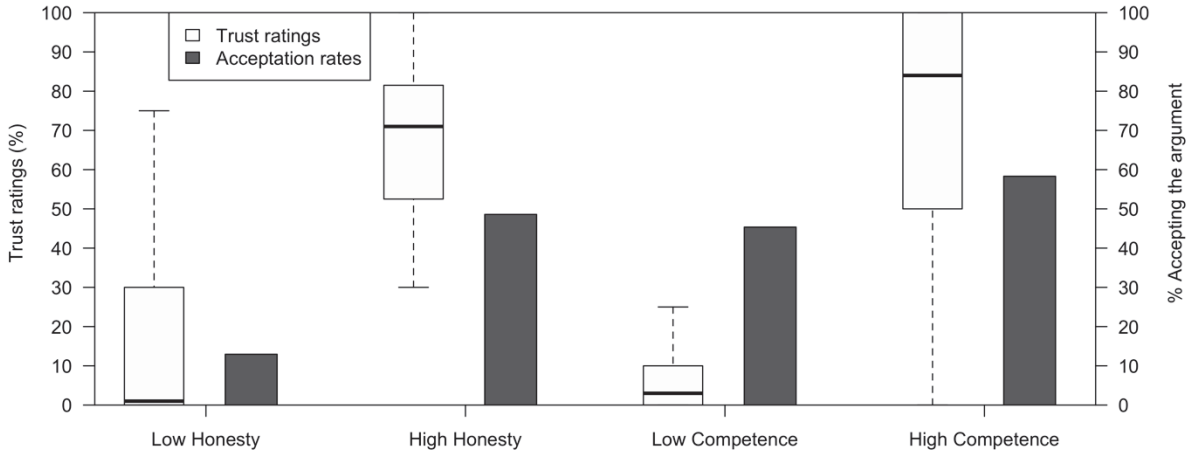


Figure 6. Trust ratings and percentage of participants accepting the argument for the correct answer in the five conditions of the trust manipulation of Experiment 4.

Discussion

As in Experiments 1 and 3, we find that being more confident in the wrong answer did not make participants more likely to reject the argument for the correct answer. This suggests that argument evaluation is not biased by how opposed participants are to the argument's conclusion. As in Experiment 2, we found that source trustworthiness had an impact on the likelihood that participants accept the argument. In the next section, we discuss two possible interpretations for the effects of source trustworthiness.

Supplementary analyses: biased immediate evaluation versus biased repetition

Experiments 1, 3, and 4 suggest that participants' confidence in the wrong answer does not affect their likelihood of accepting an argument for the correct answer. By contrast, the results regarding the influence of source trustworthiness are more ambiguous. In Experiment 3 source trustworthiness had no effect participants' likelihood of accepting the argument, but it did have some effect in Experiments 2 and 4—even though this effect did not extend to all manipulations of source trustworthiness. Here we discuss two potential explanations for these findings.

One explanation is that there is an immediate bias in argument evaluation: participants would integrate the information about source trustworthiness in their immediate evaluation of the argument, as they read it.

Another explanation is that source trustworthiness affects subsequent processes. In particular, source trustworthiness might make participants more or less likely to read the argument several times. For instance when the argument is presented by a trustworthy source, participants who immediately understand the argument should stop there, while those who fail to grasp the argument might try again. By contrast, if the source is untrustworthy, the participants who fail to grasp the argument should stop there, while those who find it persuasive might double check. We call this effect 'biased repetition.' Here we present three pieces of evidence suggesting that biased repetition accounts for the effects of trustworthiness on argument evaluation in the present experiments.

Transfer data

If argument evaluation is immediately biased by source trustworthiness, then the arguments that are accepted in part because their source is trustworthy do not need to be fully understood. To the extent that a better understanding of the argument yields increased performance at the transfer task, then the immediate biased evaluation explanation predicts that performance on the transfer task will be lower when the source was trustworthy (and the arguments only superficially examined) than when the source was untrustworthy (and the arguments had to be thoroughly understood before they were accepted).

By contrast, the biased repetition explanation predicts that arguments are only accepted once they are well understood, whether the source is trustworthy or not. This explanation thus predicts no difference in transfer rates between the trustworthy and untrustworthy sources.

In Experiment 2, among the relevant participants—i.e. the participants who adopted the correct answer after reading the argument—there were no significant differences in transfer rates (*High Trust Group*, N = 111, 77% transfer; *Low Trust Group*, N = 87, 70% transfer; $p = .26$, OR = .68). This was also true in Experiment 3 (*High Trust Group*, N = 37, 51% transfer; *Low Trust Group*, N = 23, 52% transfer; $p = 1$, *OR* = 1.03). Thus the transfer data fits better with the predictions of the biased repetition explanation than of the biased immediate evaluation explanation.

Temporal patterns

We have created simple models in order to distinguish biased repetition from biased immediate evaluation (see ESM). These models allowed us to generate differential predictions on the basis of two variables: first, how quickly the participants answer (tested here); second, how easy is the argument to understand (tested in the next section).

In Experiment 2, we analyzed the temporal pattern of the responses (Figure 7), Participants were divided in two categories—*Fast Responders* and *Slow Responders*—based on the time they took to evaluate the argument and provide a new answer to the problem (median split, with the median computed by condition). To avoid low Ns, participants were pooled into a *High Trust Group* (*High Competence* condition and *High Benevolence* condition) and a *Low Trust Group* (*Low Competence* condition and *Low Benevolence* condition). In the *High Trust Group*, *Slow Responders* (N = 71, 79%) were not significantly more likely to accept the argument than *Fast Responders* (N = 72, 76%, $p = .84$, *OR* = .87). By contrast, In the *Low Trust Group*, *Slow Responders* (N = 70, 74%) were significantly more likely to accept the argument than *Fast Responders* (N = 71, 49%, $p = .003$, *OR* = 3.0). This result fits better the predictions of the biased repetition model than those of the biased evaluation model.

Slow responders were not significantly more likely to accept the argument in the *High Trust Group* than in the *Low Trust Group* ($p = .56$, *OR* = 1.3). By contrast, *Fast responders* were significantly more likely to accept the argument in the *High Trust Group* than in the *Low Trust Group* ($p < .001$, *OR* = 3.3). This result also fits better the predictions of the biased repetition model than those of the biased evaluation model (Figure 3).

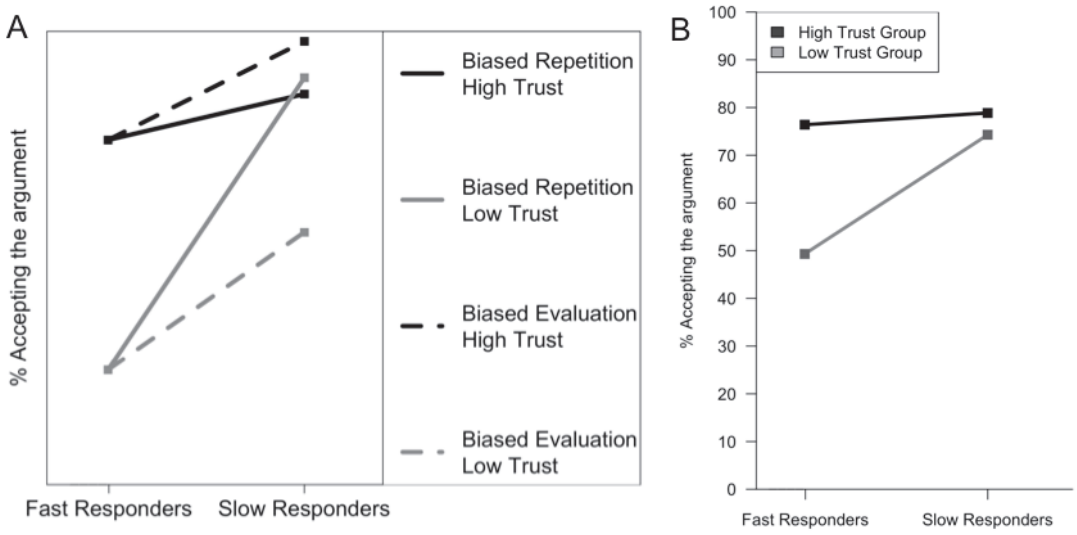


Figure 7. A) Predictions on acceptance rates as a function of reaction times and trust made by the biased repetition explanation and the biased evaluation explanation; B) Acceptance rates for Fast Responders and Slow Responders in the High Trust and Low Trust groups of Experiment 2.

Variations in difficulty of the argument

Du to the low rate of argument acceptance in the *Low Honesty* condition of Experiment 3, we could not meaningfully perform the temporal analysis performed for Experiment 2. Instead, in a follow-up study we tested predictions regarding how ease of argument understanding affects acceptance rates (see ESM). Specifically, we added a sentence to the argument explaining why the intuitive answer is incorrect (see ESM for details). We used the two conditions for which the effect on acceptance rates had been the strongest in Experiment 4: the *High Competence* condition and the *Low Honesty* condition. Although the manipulation was successful in terms of perceived trustworthiness (*High Competence* condition: N = 78, trust = 61% [*Mdn* = 72, *SD* = 36]; *Low Honesty* condition*:* N = 78, trust = 28.2% [*Mdn* = 11, *SD* = 30]), $W = 1487$, $p < .001$) there was no significant difference in the proportion of participants accepting the argument (*High Competence* condition: 43% accept; *Low Honesty* condition: N = 77, 32% accept; $p = .19$, *OR* =1.6).

Moreover, the difference between the results of this follow-up experiment and of Experiment 4 was driven by the increased acceptance rates in the low trust condition (*High Competence* condition: $p = .31$, *OR* = .63; *Low Competence* condition: $p = .02$, *OR* =3.5). As detailed in the ESM, this fits better with the pattern predicted by the biased repetition explanation than the pattern predicted by the biased immediate evaluation explanation. Finally, there were again no differences in transfer rates between conditions (high confidence *High Competence* condition: 71%, *Low Competence* condition: 71%, $p = 1$, *OR* =1.0).

Each of the three analyses—transfer rates, temporal analysis, and variation in the ease of argument understanding—support the biased repetition explanation over the biased immediate evaluation explanation.

General Discussion

Previous research suggests that argument evaluation is heavily biased, such that arguments whose conclusion challenge strongly held beliefs, or whose source is deemed untrustworthy, are evaluated particularly critically (e.g. Edwards & Smith, 1996; Hahn et al., 2009). Some results, however, suggest that a source of bias in argument

evaluation is the search for counter-arguments. When participants read arguments that challenge their views, or that come from a mistrusted source, they would rarely be swayed altogether, and they would thus generate counter-arguments. These counter-arguments would then affect the way participants evaluate the initial argument, resulting in biased *delayed* evaluation. If this is the case, then participants' *immediate* evaluation of the argument, which happens just as they read it, might not be biased.

The goal of the present experiments was to test the role of counter-arguments in argument evaluation by strongly reducing the possibility of finding counter-arguments. If participants, once they have understood the initial argument, cannot find counter-arguments, then there should be less difference between their immediate evaluation of the argument and their delayed evaluation, the one that is recorded by the experimenters.

We used arguments that, once they are understood, do not suffer any good counter-arguments. If it is the case that it is largely counter-arguments that drive bias in argument evaluation, then argument evaluation should be largely free of bias in this case. This would mean that participants are equally likely to understand, and then to accept, arguments that challenge their views or that come from untrustworthy sources than arguments that are less challenging or that come from trusted sources.

In Experiments 1, 3, and 4, we manipulated the degree to which the argument's conclusion challenged the participants' views. To do so, we induced participants to be more or less confident in their answer, answer that was then challenged by the argument. In all three experiments the manipulation was successful, so that participants in one condition had more confidence in their wrong answer than the participants in the other condition. Yet being more confident in the wrong answer did not make participants less likely to accept the argument for the correct answer.

In Experiments 2, 3, and 4, we manipulated the degree to which the argument's source was trustworthy. To do so, we led participants to believe that the arguments' source was either competent, incompetent, honest, or dishonest. In all three experiments the manipulation was successful, so that participants in the high trust conditions (competent or honest sources) rated the source as much more trustworthy than participants in the low trust conditions (incompetent or dishonest sources). In Experiment 3, arguments from trustworthy sources were not more likely to be accepted than arguments from untrustworthy sources. By contrast, in Experiments 2 and 4, some forms of (un)trustworthiness made participants more or less likely to accept the argument.

In a series of supplementary analyses, we offered evidence suggesting that these results can be interpreted in terms of biased repetition. Participants would not be immediately biased when they read the argument. Instead, they would be more or less likely to read the argument again as a function of the source's trustworthiness and the outcome of their first reading of the argument. For instance, a participant who failed to understand the argument would be more likely to read it again if the argument's source is trustworthy than untrustworthy.

On the whole, the results from the four present experiments suggest that when participants cannot easily generate counter-arguments, then they evaluate arguments in a way that is largely unbiased: they are equally likely to understand and accept arguments that are more challenging than arguments that are less challenging, and arguments that come from untrustworthy sources than trustworthy sources.

In combination with the other results demonstrating the role of counter-arguments in biased argument evaluation, our results thus suggest that immediate argument evaluation, the evaluation which takes place as people read or hear an argument, could be largely unbiased. If correct, this conclusion would have significant methodological, theoretical, and practical consequences.

From a methodological point of view, attention should be drawn to the distinction between immediate and delayed argument evaluation. The cognitive mechanisms at play are likely different when people immediately evaluate an argument as they read or hear it and when they take their time to think of an explicit argument rating. In particular, the cognitive mechanisms tasked with evaluating arguments would be more directly studied by examining immediate argument evaluation. Delayed argument evaluation, by contrast, likely involves several other cognitive processes, in particular argument production (required for the production of counter-arguments). In order to study the cognitive mechanisms underpinning argument evaluation, it would thus be desirable to adopt techniques that can immediate argument evaluation instead of delayed argument evaluation.

The degree to which immediate argument evaluation is biased is also of theoretical import. In particular, it is relevant for some predictions made by the argumentative theory of reasoning (Mercier & Sperber, 2011). According to this theory, the main function of reasoning is to argue: to produce arguments in order to convince others, and to evaluate others' arguments in order to know whether we should be convinced. The argumentative theory predicts that reasoning should have a my-side bias (or confirmation bias) when it produces arguments, since arguments that support our point of view are more likely to convince an audience (Mercier, in press). By contrast the chief goal of argument evaluation would be to recognize good enough arguments so that one can change one's mind when warranted. If arguments that challenge our views or come from mistrusted sources were too easily rejected, argumentation would be pointless. As a result, argument evaluation should be as objective as possible.

Findings of biases in argument evaluation thus speak against the argumentative theory of reasoning. However, if the explanation suggested here is correct, then these results might in fact be entirely consistent with the theory. Participants would not be biased in their immediate evaluation, but they would be bias when they subsequently start to produce counter-arguments, as expected by the theory.

Finally, the question of whether biases in argument evaluation arise immediately or only after the generation of counter-arguments also has practical consequences. If the biases were present immediately, as people read or hear arguments, they would be very difficult to counteract. By contrast, the effects of the generation of counter-arguments can be more easily remedied by addressing them. If the counter-arguments are addressed, then we should expect people to fall back on their initial, unbiased evaluation of the argument. By contrast, if argument evaluation were really biased, then seeing their counter-arguments addressed should not stop people from rejecting the initial argument.

Thus the outcome of group discussion, to the extent that the discussion successfully addresses counter-arguments, can also provide a test of how biased argument evaluation really is. In previous research using the same problems as in the present experiments, we observed that when participants were allowed to discuss the problems in small groups, then the participant(s) with the correct answer was always able to convince the group that she was correct (or nearly always in the case of 10-year-olds, Trouche et al., 2014). This was true even when this participant was alone, and even if she was less confident than the other group members. Thus, when a discussion allows all potential counter-arguments and misunderstandings to be addressed, argument strength trumped confidence, a potentially biasing factor.

Other relevant evidence had been gathered in the framework of Persuasive Argument Theory (Vinokur, 1971). According to this theory, the outcome of group discussions can be explained by the quantity and quality of the arguments exchanged. Source factors, such as whether an opinion is defended by a majority or a minority of the group members, only play an indirect role, through the number of arguments generated for each opinion. The

analyses of different types of group discussion support the Persuasive Argument Theory (for review, see, Isenberg, 1986).

More generally, group discussions allow individuals to change their mind—generally for the best—in a wide variety of contexts (for reviews, see, Laughlin, 2011; Mercier, 2011a, 2011b; Mercier & Sperber, 2011). If argument evaluation were heavily biased, people should not be able to change their mind and adopt better-supported opinions so efficiently. The outcome of group discussion is more consistent with an unbiased ability to evaluate arguments.

To conclude, we would like to stress that believing that argument evaluation is irremediably biased might have negative practical consequences. For instance, it might explain why people vastly underestimate the benefits of group discussion (Mercier, Trouche, Yama, Heintz, & Girotto, in press). This underestimation, in turn, might preclude people from taking full advantage of the benefits of argumentation. If other people are thought to be so biased that they reject arguments as soon as they disagree with the arguments' conclusion, then there is not much point engaging in discussion. Fortunately, this seems to be wrong: people might be able to evaluate arguments in an unbiased manner—and group discussion is the context in which this skill is most likely to be expressed.

References

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*(1), 47.

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little, Brown.

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., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Greenwald, A. G. (1968). Cognitive learning, cognitive response to persuasion, and attitude change'. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological Foundations of Attitudes* (pp. 147–170). New York: Academic Press.

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337–367.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732.

Hoeken, H., Šorm, E., & Schellens, P. J. (2014). Arguing about the likelihood of consequences: Laypeople's criteria to distinguish strong arguments from weak ones. *Thinking & Reasoning*, *20*(1), 77–98.

Hornikx, J., & Hahn, U. (2012). Reasoning and argumentation: Towards an integrated psychology of argumentation. *Thinking & Reasoning*, *18*(3), 225–243.

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141–1151.

Klaczynski, P. A., & Gordon, D. H. (1996). Self-serving influences on adolescents' evaluations of belief-relevant evidence. *Journal of Experimental Child Psychology*, *62*, 317–339.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory and Cognition*, *6*, 107–118.

Kuhn, D., & Lao, J. (1996). Effects of Evidence on Attitudes: Is Polarization the Norm? *Psychological Science*, *7*, 115–120.

Laughlin, P. R. (2011). *Group problem solving*. Princeton: Princeton University Press.

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence*, *30*(1), 81–108.

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.

Mercier, H. (in press). Confirmation (or myside) bias. In R. Pohl (Ed.), *Cognitive Illusions* (2nd ed.). London: Psychology Press.

Mercier, H. (submitted). How gullible are we? A review of the evidence from psychology and social science.

Mercier, H. (2011a). Reasoning serves argumentation in children. *Cognitive Development*, *26*(3), 177–191.

Mercier, H. (2011b). When experts argue: explaining the best and the worst of reasoning. *Argumentation*, *25*(3), 313–327.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Miyamoto, Y., Nisbett, R. E., & Masuda, T. (2006). Culture and the physical environment: Holistic versus analytic perceptual affordances. *Psychological Science*, *16*(2), 113–119.

Munro, G. D., Ditto, P. H., Lockhart, L. K., Fagerlin, A., Gready, M., & Peterson, E. (2002). Biased assimilation of sociopolitical arguments: Evaluating the 1996 US presidential debate. *Basic and Applied Social Psychology*, *24*(1), 15–26.

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Science*, *9*, 467–473.

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*(5), 653–684.

Petty, R. E., & Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. In D. T. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 323–390). Boston: McGraw-Hill.

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.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, *25*(4), 359–393.

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. (1978). Self-generated attitude change. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 289–338). New York: Academic Press.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

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.

SUPPLEMENTARY MATERIAL

Experiment 1

Previous research has shown that this problem has the properties required to test how biased is argument evaluation when the argument suffers no good counter-argument (Trouche, Sander, & Mercier, in press). Most participants provide the intuitive wrong answer ("Cannot be determined") and, among those participants, about half change their mind when confronted with an argument for the correct answer (Trouche et al., in press). These results thus avoid ceiling or floor effects when manipulating factors supposed to affect the likelihood that the argument is accepted.

Confidence scale

The numbers in brackets were presented to the participants.

0 - I answered randomly (1)

1 - Not confident at all (2)

2 - A little confident (3)

3 - Somewhat confident (4)

4 - Quite confident (5)

5 - Very confident (6)

6 - Extremely confident (7)

7 - As confident as the things I'm most confident about (8)

Transfer problems

Second problem

Paul is selling bonds to Liz and Liz is selling bonds to Steven.
Paul works for a company called GNB and Steven also works for GNB.
Is there someone working for GNB selling bonds to someone who doesn't work for GNB?

Third problem

Jeff is selling bonds to Susan and Susan is selling bonds to David.
Jeff works for a company called GNB, and David doesn't work for GNB .
Is there someone working for GNB selling bonds to someone who doesn't work for GNB?

Manipulation check data for all the participants

(Low Confidence, N = 80, confidence = 67.3% [Mdn = 71.4, SD = 26.1]; High Confidence, N = 68, confidence = 80.3% [Mdn = 85.7, SD = 20.5]; W = 1943, p = .002)

Number of excluded participants in the analyses of % of participants accepting the correct answer

None

Transfer rates

When looking at participants who accepted the argument for the good answer in both conditions, transfer rates did not differ significantly between the two conditions.

High Confidence condition, N = 36, 33% transfer; Low Confidence condition, N = 29, 48% transfer; p = .31, OR = .54).

Experiment 2

Manipulation check data for all the participants

Trust was higher in the High Honesty condition (N = 101, trust = 71.4% [Mdn = 70.0, SD = 17.4]) than in the Control condition (N = 99, trust = 56.2% [Mdn = 52.0, SD = 21.4]; W = 3687, p = .001), and the Low Honesty condition (N = 99, trust = 17.7% [Mdn = 10.0, SD = 24.1]; W = 776, p < .001).

It was also higher in the Control Condition than in the Low Honesty condition (W = 1304, p < .001). Trust was higher in the High Competence condition (N = 101, trust = 80.9% [Mdn = 86.0, SD = 18.1]) than in the Control condition (W = 1860, p < .001), and the Low Competence condition (N = 100, trust = 14.4% [Mdn = 2.5, SD = 23.2]; W = 388, p < .001).

It was also higher in the Control condition than in the Low Competence condition (W = 1122, p < .001)

Number of excluded participants in the analyses of % of participants accepting the correct answer

For the conditions High Honesty, Low Honesty, High Competence and Low Competence, 7 subject were excluded (respectively 2, 2, 2, and 1) because they did change their answer but for the other wrong answer ("No").

Pilot for Experiment 3

Participants

57 participants were recruited online through the crowdsourcing website Sojump (39 females, Mage = 30.5, SD = 4.8). They were paid 1.5 Yuan for their participation and had to be located in the People's Republic of China.

Materials and procedure

Participants were asked to solve the disjunctive reasoning problem described in the main text, to provide a justification for their answer, and to evaluate their confidence in the correctness of their answer. They were then told that they would be given some answers and the arguments given by other participants. Two different answers were presented along with their justifications, one for the intuitive wrong answer and the other for the correct answer. The two conditions differ only by the display order of the two arguments (Right first or Wrong first). Finally, the same two transfer problems were asked as in the previous experiments.

Results

As for the American population, the large majority of participants (70%) gave the intuitive but wrong answer. For those 40 participants, the acceptation rates did not differ significantly between the two conditions

(Right First: 57%, Wrong First: 74%, p = .33, OR =.49). Looking at participants who accept the argument only, transfer rates in each condition were respectively of 50% (N = 12) and 57% (N = 14).

Experiment 3

Manipulation check data for all the participants

Confidence

(Low Confidence, N = 80, confidence = 68.6% [Mdn = 71.4, SD = 29.0]; High Confidence, N = 77, confidence = 86.3% [Mdn = 100.0, SD = 19.7]; W = 1901, p < .001).

Trust

Trust was higher in the High Honesty condition (N = 76, trust = 71.7% [Mdn = 75.0, SD = 21.2]) than in the Low Honesty condition (N = 75, trust = 42.1% [Mdn = 43.0, SD = 29.6]), W = 1205, p < .001). Trust was higher in the High Competence condition (N = 74, trust = 73.2% [Mdn = 80.0, SD = 32.1]) than in the Low Competence condition (N = 80, trust = 41.4% [Mdn = 44.0, SD = 32.1]), W = 1306, p < .001)

Number of excluded participants in the analyses of % of participants accepting the correct answer

For the conditions High Honesty, Low Honesty, High Competence and Low Competence, 3 subject were excluded (respectively 0, 1, 1 and 1) because they did changed their answer but for the other wrong answer ("No").

Transfer data

Among participants who accepted the argument for the good answer in all conditions, transfer rates did not differ significantly between the conditions.

Control condition, N = 49, 53% transfer;

High Honesty condition, N = 61, 36% transfer;

Low Honesty condition, N = 43, 49% transfer;

High Competence condition, N = 60, 58% transfer;

Low Competence condition, N = 44, 50% transfer;

Comparing the largest difference in term of transfer rate (High Honesty condition versus High Competence condition) leads to no significant difference (p = .13, OR = .54).

Pilot Experiment 4

50 participants were recruited online through through Amazon Mechanical Turk (24 females, Mage = 32.6, SD = 11.4). They were paid $0.70 for their participation and had to be located in the U.S.

Materials and procedure

Participants were asked to solve the problem described in the main text, to provide a justification for their answer, and to evaluate their confidence in the correctness of their answer. Then after being told that some answers and arguments given by other participants would be given to them, two different answers were presented along with their justifications, one for the intuitive wrong answer and the other for the correct answer.

The two conditions differ only by the display order of the two arguments (Right first or Wrong first). Finally, we asked participants whether they have seen this problem before or not, along with a "not sure" option.

Results

26 (52%) participants gave the intuitive but wrong answer, surprisingly differing between the two conditions (18 (69%) for the Right First condition and 8 (33%) for the Wrong First condition). Among those participants, acceptation rates were however comparable, respectively 44% and 25%.

Experiment 4

Trust

Trust was higher in the High Honesty condition (N = 76, trust = 67.6% [Mdn = 70.0, SD = 20.8]) than in the Low Honesty condition (N = 78, trust = 21.2% [Mdn = 9.0, SD = 27.1]), W = 682, $p < .001$). Trust was higher in the High Competence condition (N = 67, trust = 71.8% [Mdn = 84.0, SD = 33.5]) than in the Low Competence condition (N = 74, trust = 8.35% [Mdn = 2.0, SD = 14.6]), W = 488, $p < .001$)

Number of excluded participants in the analyses of % of participants accepting the correct answer

Only 2 subjects, both in the Low Honesty condition, were excluded because they did changed their answer but for the other wrong answer ("6").

Transfer problem:

Jeff buys a CD and a book, he has to pay 16$. The CD costs 10$. How much is the book?

A bag of concrete weighs 22 ounces. The concrete weighs 20 ounces more than the bag. How much does the bag weigh?

Transfer Data:

Looking at the transfer rates of participants who accepted the argument for the good answer en each condition:

High Honesty condition, N = 18, 66% transfer;

Low Honesty condition, N = 19, 37% transfer;

High Competence condition, N = 5, 60% transfer;

Low Competence condition, N = 18, 50% transfer;

Models

We created simple models to make predictions based on two different explanations of potential effects of trust in the argument's source we might obtain. The first explanation is biased repetition, in which truth influences the chances that participants read the argument again, but not how they evaluate the argument each time they read it. The second explanation is biased evaluation, in which trust directly influences how participants evaluate the argument.

To model these two explanations, we used three parameters: the probability that the argument is understood each time it is read (pU), the probability that it is read again if it was previously understood (pRU), and the probability that it is read again if it was previously not understood (pRNU). If argument evaluation is unbiased, but biased repetition is at play, then increased source trustworthiness should: have no effect on pU, increase pRNU, and decrease pRU. By contrast, if argument evaluation is biased, then source trustworthiness should directly influence pU.

On the basis of these parameters, and starting with a population of N participants, it is possible to count how many participants answer that they accept or reject the argument at each time step (i.e. after each reading of the argument). For the biased repetition explanation, we used the following parameter values:

High Trust: pU(0.5); pRU(0.25); pRNU(0.75).

Low Trust: pU(0.5); pRU(0.75); pRNU(0.25).

For the biased evaluation explanation, we used the following three sets of parameter values:

High Trust: pU(0.75); pRU(0.5); pRNU(0.5).

Low Trust: pU(0.25); pRU(0.5); pRNU(0.5).

The simulations based on these parameter values can be found at https://sites.google.com/site/hugomercier/simulations%20unbiased%20argument%20evaluation.xlsx?attredirects=0. To simulate the temporal pattern predicted by the biased repetition and the biased evaluation explanations, we compared the proportion of participants immediately accepting the argument to the proportion of participants accepting the argument after at least one repetition. The predictions of the biased repetition model are as follows (see Figure 7 of the main text—note that these predictions are not meant to be quantitative, which is why we have removed the scales from the prediction figures). When the source is trustworthy, participants who answer quickly should be as likely to accept the argument as those who answer slowly. By contrast, when the source is untrustworthy, participants who answer quickly should be much less likely to accept the argument than those who answer slowly. Moreover, source trustworthiness should have little effect among the participants who answer slowly. By contrast, if biased evaluation is at play, then source trustworthiness should have much less influence on the temporal pattern of the responses: there should be large differences between the acceptance rates for trustworthy and untrustworthy sources both among fast and slow responders.

To simulate the effect of variations in the ease with which the argument is understood, we compared the proportion of participants who end up accepting the argument while varying pU (the probability that the argument is understood). For the biased repetition explanation, we used the following parameter values:

Easy Argument / High Trust: pU(0.25); pRU(0.1); pRNU(0.9).

Easy Argument / Low Trust: pU(0.75); pRU(0.9); pRNU(0.1).

Hard Argument / High Trust: pU(0.25); pRU(0.1); pRNU(0.9).

Hard Argument / Low Trust: pU(0.75); pRU(0.9); pRNU(0.1).

For the biased evaluation explanation, we used the following parameter values:

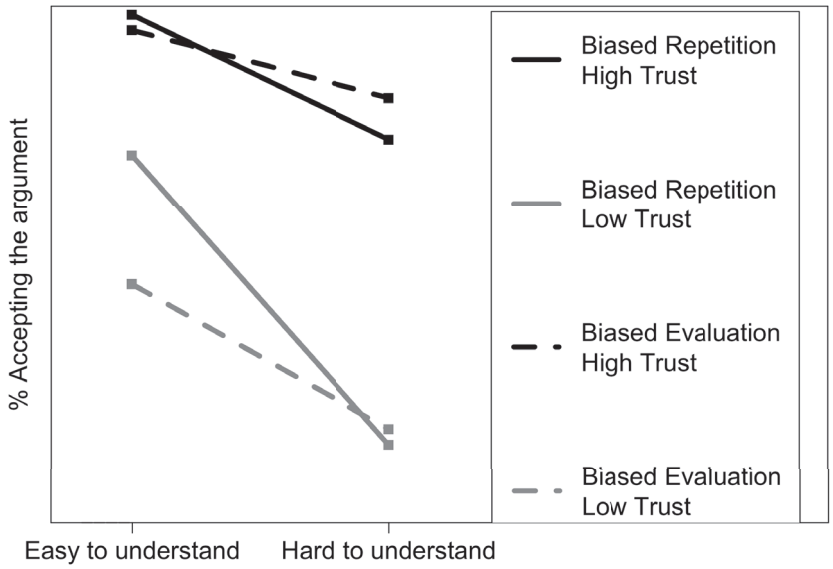
Easy Argument / High Trust: pU(0.9); pRU(0.5); pRNU(0.5).

Easy Argument / Low Trust: pU(0.6); pRU(0.5); pRNU(0.5).

Hard Argument / High Trust: pU(0.4); pRU(0.5); pRNU(0.5).

Hard Argument / Low Trust: pU(0.1); pRU(0.5); pRNU(0.5).

The outcome of the simulation was as follows:



If the biased repetition explanation is correct, we should observe a stronger interaction between trustworthiness and ease of understanding. In particular, when the arguments are easy to understand, source trustworthiness should make relatively little difference, while it should have a large influence when the arguments are hard to understand. By contrast, if the biased evaluation explanation is correct, then we expect the effects of ease of understanding and of source trustworthiness to be mostly additive, so that even if the arguments are easy to understand, we should still observe a large gap between the acceptance rates for trustworthy and untrustworthy sources.

Experiment with easy to understand arguments

299 participants were recruited online through Amazon Mechanical Turk (132 females, Mage = 34.6, SD = 11.3). They were paid $0.70 for their participation and had to be located in the U.S.

Materials and procedure

The materials and procedure were exactly identical to the High Competence and Low Benevolence conditions of Experiment 4, except the argument used. In experiment 4 it was:

“It says that there are 8 more women than men, so if there are 2 men, there are 2 + 8 women, so 10 women. And 2 men plus 10 women makes 12 people jury. »

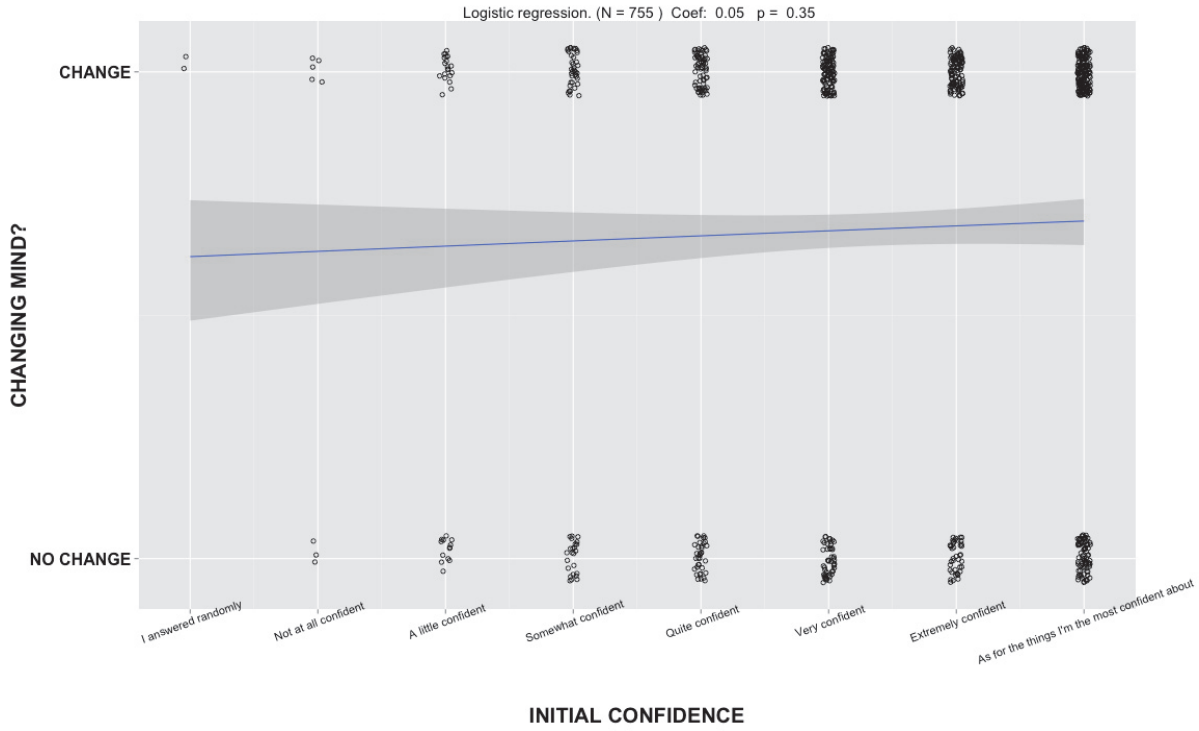
Now it is:

“It can't be 4, because if there were 4 men, then there would be 12 women (8 more than the men), and that would make for a total of 16 people. 2 works. If there are 2 men, then there are 10 women, and a total of 12.”

Results (for all participants)

Trust was significantly higher in the High Competence condition (N = 151, trust = 67.2% [Mdn = 81.0, SD = 34.5]) than in the Low Honesty condition (N = 148, trust = 22.0% [Mdn = 10, SD = 26.8]), W = 3923, $p < .001$).

A l'image du film « 12 hommes en colère », l'échange de raisons peut surmonter la sur-confiance, le manque de confiance en la source et même l'infériorité numérique. Dans l'article (Trouche, Sander & Mercier 2014), nous avons observé, chez des élèves de CM2, des cas où, dans un groupe de 4, l'élève qui avait la bonne réponse faisait face à 3 autres élèves plus confiants que lui et défendant tous la même réponse intuitive. Comme étudié dans cet article non intégré à la thèse, dans la majorité de ces cas, la bonne réponse l'emportait grâce à l'échange d'arguments. Ajoutons un résultat sur le sentiment de confiance allant dans le même sens à partir d'un agrégat de nos données recueillies en ligne sur la tâche de Paul et Linda. La confiance initiale des sujets ayant la réponse intuitive avant de recevoir un argument pour la bonne réponse n'est pas un bon prédicteur de : qui va changer d'avis face à l'argument ? Le graphique ci-dessous regroupe 755 sujets et montre qu'entre les sujets qui changent d'avis et les autres, la confiance initiale est également distribuée.



Cela illustre la trajectoire relativement typique du sentiment de confiance des sujets qui changent d'avis dans nos expériences, exprimant une confiance en leur réponse très haute, jusqu'à ce qu'ils changent d'avis face à des arguments acceptés pour ensuite afficher une confiance à nouveau très haute pour leur nouvelle réponse.

Nous avons jusque là supposé que c'était l'échange de raisons qui était responsable des réussites en groupe, mais est-ce la seule explication possible ? On pourrait par exemple penser que le simple fait de se retrouver en groupe motiverait plus les sujets que ne le fait le raisonnement en solitaire. Les améliorations observées des discussions de groupe ne proviendraient que d'une motivation supérieure des individus. Cependant, même lorsque les sujets sont motivés financièrement, cela n'améliore, en général, peu ou pas du tout leurs performances individuelles. Par exemple, Johnson-Laird & Byrne 2002 rapportent une expérience réalisée par Santamaria et Johnson-Laird où les chercheurs avaient promis une récompense financière aux sujets qui trouveraient la bonne réponse à la tâche de Wason. Ils n'observèrent aucune amélioration.

On pourrait également imaginer que le fait de savoir que d'autres personnes ont donné une réponse différente de la leur suffirait à ce que les gens réalisent leurs erreurs. C'est ce que nous avons testé dans la dernière étude expérimentale présentée dans cette thèse.

Les 226 étudiants en Licence de Sciences cognitives étaient répartis en 12 groupes. Les groupes étaient séparés en deux conditions. Une condition « discussion » similaire au paradigme de diffusion présenté précédemment où les sujets ont comme consigne de se mettre d'accord avec tous leurs voisins directs, indiquant leur réponse et leur sentiment de confiance toute les minutes. Dans l'autre condition, la condition « silence », les sujets peuvent voir les réponses de tous leurs voisins directs, mais ne peuvent pas discuter entre eux. Chaque groupe d'étudiant, composé de 11 à 25 membres, est soit dans la condition Silence, soit dans la condition Discussion. Il doivent d'abord résoudre individuellement le problème du bonbon et de la baguette, ou celui de Paul et Linda, puis passent à la phase sociale du premier problème. Enfin, tous les groupes recommencent la même séquence phase individuelle puis phase sociale avec l'autre problème (l'ordre était contrebalancé).

Cela correspond à la première étude de l'article ci-dessous, dont nous discuterons les résultats ensuite. Nous ne discuterons cependant pas la deuxième étude de l'article portant sur les chaines de transmission, dépassant le cadre de cette thèse.

Cet étude est le fruit d'une collaboration avec Nicolas Claidière et n'a pas encore fait l'objet d'une publication.

Argumentation and the diffusion of counter-intuitive beliefs

Nicolas Claidière
LPC, CNRS, University Aix-Marseille

Emmanuel Trouche
L2C2, CNRS, University Lyon 1

Hugo Mercier
University of Neuchâtel

Abstract

Research in cultural evolution has focused on the spread of intuitive or minimally counter-intuitive representations. However, some very counter-intuitive representations can also spread successfully, at least in some communities—scientific theories being the most prominent example. We suggest that argumentation could be an important factor in the spread of counter-intuitive representations. A first experiment demonstrates that argumentation enables the spread of the counter-intuitive answer to a reasoning problem in large discussion groups, whereas this spread is limited or absent when participants can show their answers to each other but cannot discuss. A series of experiments using the technique of repeated transmission show that, in the case of the counter-intuitive representation studied: a) arguments can help spread this representation without loss; b) conformity bias does not help spread this representations; c) prestige bias plays a minimal role in helping spread this representation. Argumentation thus seems to be necessary and sufficient for the spread of some counter-intuitive representations.

Keywords

Counter-intuitive representations; argumentation; cultural evolution; conformity bias; prestige bias.

Introduction

Some ideas have managed to spread in human societies despite being highly counter-intuitive, from heliocentrism to the Christian Trinity. Remarkably, these ideas have spread in the face of beliefs that were not only more intuitive, but also more widespread and held by the most prestigious members of the relevant group, raising an interesting challenge for the study of cultural evolution.

Studies have shown that cultural evolution often converges on the most intuitive cultural variant—the variant that triggers the most inferences for the least effort—whether it is in humans (language: Griffiths, Kalish, & Lewandowsky, 2008; Kirby, Cornish, & Smith, 2008; Reali & Griffiths, 2009; medicine: Miton, Claidière, & Mercier, 2015; art: Morin, 2013) or in other animals (visual stimuli: Claidière, Smith, Kirby, & Fagot, 2014; bird song: Feher, Wang, Saar, Mitra, & Tchernichovski, 2009; foraging strategy: Laland & Williams, 1997). Theoretical analyses have also revealed the origin and strength of this result: when cultural variants spread in a population, transformations that occur during transmission progressively accumulate and tend to be directed by pre-existing biases (Claidière & Sperber, 2010; Griffiths et al., 2008; Kalish, Griffiths, & Lewandowsky, 2007; Kirby, Dowman, & Griffiths, 2007). Given the strength and generality of these results it is intriguing that counter-intuitive beliefs can sometimes dislodge and replace more intuitive variants.

One possible explanation relies on the properties of counter-intuitive beliefs: counter-intuitive ideas can become attractive in virtue of their counter-intuitiveness. For instance, many beliefs in religious and other supernatural entities are counter-intuitive: a ghost has the counter-intuitive property of being invisible. Boyer (2001; see also Sperber, 1996) has suggested that in fact such beliefs are ideal for cultural transmission because they are *minimally* counter-intuitive: a ghost is invisible but has the mind of a human being (so we intuitively understand its motives for instance) and is therefore both easy to understand (since it is mostly intuitive), and appealing (because of the counter-intuitive property). According to this explanation, the cultural evolution of minimally counter-intuitive beliefs follows from the general principles highlighted previously: the more appealing beliefs tend to spread. However, many beliefs that are much more than minimally counter-intuitive have spread and remained stable (the revolution of the earth around the sun for instance).

A possible explanation of the spread of (non-minimally) counter-intuitive beliefs relies not on the intrinsic properties of cultural variants but on their source. For instance, if a prestigious individual adopts a counter-intuitive belief, this belief could then be adopted by other members of the population who imitate prestigious individuals (Boyd & Richerson, 1985; Richerson & Boyd, 2005). The spread of this belief could then be reinforced and stabilized by a conformist tendency (i.e. the adoption of beliefs held by a majority of individuals). Boyd and Richerson in particular have argued that such processes can lead in

some cases to the spread and stabilization of any cultural variants, including maladaptive or counter-intuitive ones. One could thus imagine that a combination of prestige bias and conformity bias could account for the spread and stability of counter-intuitive beliefs. It is indeed likely that these factors play an important role in the adoption of counter-intuitive beliefs by most people. For instance, most people nowadays believe in scientific theories through trust in teachers and scientists, and (maybe) conformity.

A potential difficulty with this explanation is that some counter-intuitive beliefs started spreading in spite of being defended by individuals who were not particularly prestigious. On the contrary, it is only after the counter-intuitive beliefs had been accepted in the community and their value recognized that their creators were endowed with prestige. Einstein, Galileo, or Newton are good examples, but that is true of just about any influential scientist, and also, to some extent, of theologians and philosophers, who also spread counter-intuitive beliefs.

We propose that argumentation plays a crucial role in the propagation of such counter intuitive beliefs and gives rise to a hitherto unrecognized evolutionary dynamic. In order to study in the laboratory the effects of argumentation on the spread of counter-intuitive beliefs, we can rely on reasoning problems that have an intuitive but wrong answer, given by most participants, and a counter-intuitive but correct answer that only a few participants reach. For instance, consider the bat and ball problem (Frederick, 2005):

A bat and a ball cost $1.10 together. The bat costs $1 more than the ball. How much does the ball cost?

Studies have shown that a majority of participants provide the intuitive but incorrect answer of 10c, when the correct but counter-intuitive answer is 5c (5c for the ball plus $1.05 for the bat makes $1.10 in total).

For such problems in which the correct answer logically follows from knowledge the participants agree on a participant who defends the correct answer typically convinces the members of a small group (e.g. $N = 4$) to accept it (Laughlin, 2011; Trouche, Sander, & Mercier, 2014). At least two elements suggest that argumentation plays a crucial role in the transmission of the correct but counter-intuitive answer in these small groups. First, the transcripts show participants exchanging arguments and being convinced only when good arguments are offered (Moshman & Geil, 1998; Trognon, 1993). Second, other factors such as support from other individuals or confidence seem to play a minimal role since the correct response spreads even when it is defended by a single individual facing a unanimous group, and even if this individual is less confident than the other members (Trouche et al., 2014).

If these findings suggest that argumentation can spread counter-intuitive beliefs, several questions remain unaddressed. Firstly, to demonstrate that argumentation can spread beliefs in a large population it is necessary to show that participants that have been convinced to adopt the correct answer can themselves convince others who can convince others, and so on. Previous work on cultural transmission

showed the progressive erosion of information in various tasks (e.g. Bartlett, 1932; Maxwell, 1936; Mesoudi & Whiten, 2004; Northway, 1936; Scott-Phillips, in press). This erosion may act against the spread of arguments and one might predict that across several generation of transmission arguments become less and less elaborate and therefore less and less convincing. For instance, the arguments (along with the stories, descriptions, pictures, and prose) used by Bartlett lost nearly all of their content after a few transmission episodes (Bartlett, 1932). Secondly, a counter-intuitive belief spreads in replacement of a more common belief and therefore has to overcome source-based biases such as prestige and conformity biases.

The following experiments seek to establish: a) In study 1, that argumentation can enable the spread of counter-intuitive beliefs in large groups and across many generations without erosion; b) In study 2a and 2b, that being exposed to a single argument, instead of a full-blown argumentative discussion, can allow the spread of counter-intuitive beliefs; c) In study 3a, b and c, that other factors related to the source of the belief—how many people hold it, and how prestigious is the source holding it—are less efficient than argumentation in spreading counter-intuitive beliefs.

Materials for all studies

In all studies we rely on two problems, the bat and ball, described above and Paul and Linda:

Paul and Linda problem

Paul is looking at Linda and Linda is looking at John.

Paul is married but John is not married.

Is a person who is married looking at a person who is not married?

Yes, someone who is married is looking at someone who is not married

No, no one who is married is looking at someone who is not married

Cannot be determined whether someone who is married is looking at someone who is not married

In this problem as well, the majority of participants provide the incorrect answer "Cannot be determined," when the correct but counter-intuitive answer is "Yes, someone who is married is looking at someone who is not married" (since Linda has to be either married or not married, and that the statement is true in both cases) (Toplak & Stanovich, 2002).

The bat and ball and Paul and Linda are intellective problems (see Laughlin & Ellis, 1986) in the sense that the participants can understand the correct answer on the basis of the information provided and their prior understanding of mathematics (Bat and Ball) and logic (Paul and Linda). We chose these problems because they are well studied reasoning problems that have been shown to have an intuitive but wrong answer given by most participants.

Study 1: diffusion of counter-intuitive beliefs in large groups through discussion

These experiments extended to larger groups the previous studies showing that argumentation enables the diffusion of counter-intuitive beliefs in small groups. For counter-intuitive beliefs to spread in large groups, participants who did not find the correct answer on their own, but who have been convinced to accept it must be able to convince others in turn. To establish that this is the case, we tested

much larger groups than the groups usually tested (mean group size=18.8 participants). Moreover, we kept track of the potential diffusion of the answers by asking participants to provide answers at regular time intervals throughout the experiment.

In order to show that it is argumentation that explains the spread of counter-intuitive beliefs we use, as control condition, groups in which participants could only show their response to others but could not discuss them (see Minson, Liberman, & Ross, 2011; Rahwan, Krasnoshtan, Shariff, & Bonnefon, 2014; Rowe & Wright, 1996).

Participants

226 participants (first year students at the University of Lyon) were recruited (71 females, M_{Age} = 19.4, *SD* = 2.1). They were distributed based on their class assignment to 12 groups of varying sizes (*Max* = 25, *Min* = 11, *Mean* = 19). This was the first course of the year, for first year psychology students, so we can assume that most students did not know each other before the experiment.

Materials

Participants completed the Bat and Ball problem and the Paul and Linda problem in counterbalanced order. For the Bat and Ball, the answer format was open ended. For the Paul and Linda problem, the participants had to choose one of the three possible answers. In both cases participants had to indicate their confidence in their answer on a confidence scale going from 0 to 10 (the results from this question will not be presented in detail here, but the data are available in the ESM).

Design and procedure

Six groups took part in the Discussion condition, and six in the Silence condition. Both conditions had two phases: an Individual phase and a Social phase. The Individual phase, presently described, was identical across the two conditions.

Individual phase. After agreeing to take part in the experiment, the participants were made to sit so that the seating arrangement could best approximate a rectangle with no empty seats. Answer sheets were distributed which contained 25 identical rows with the space for an answer to the problem and a confidence scale. After a brief explanation of the first phase of the experiment, the experiment started. The problem was displayed on the screen so that all participants could start completing it at the same time. After 20 seconds, the participants provided their first answer and confidence rating. More answers and confidence ratings were gathered at one-minute intervals four times.

Social phase. In the Discussion condition, participants were told that they would now be able to discuss their answers with their neighbors. Neighbors were defined as the eight (maximum) students surrounding them. Participants were told "The goal is to reach a consensus for the whole group. So after you have made sure that you agreed with some of your neighbors, you should turn to your other neighbors to make sure that they also agree on the same answer." After they were given the signal to start discussing, the participants had to write down their answer and confidence rating every minute. After 5 minutes the participants were asked every other minute if at least one of them had changed their mind. The experiment was stopped when no one had changed their mind. For this reason the length of the

experiment varied both within and between conditions (from 8 to 23 measures). Time was kept by the experimenter who required everyone to write down their answers every minute. The instructions were identical in the Silence condition, except that participants were instructed only to look at the answers of their neighbors, and were prohibited from talking or writing anything besides their answers.

Statistics

We analyzed the results using Generalized Linear Mixed Models (GLMM) and followed the procedure recommended by Zuur *et al.* (2009). The dependent variable was the success of each participant at each time step (binary variable). We included the Group as a random variable with a random intercept and a random slope depending on time to account for repeated measurements. In order to compare the results between the individual and the group phase we limit our analysis to five measures in each phase.

Based on the design of the experiments we chose to include three explanatory variables. The first variable represented the phase of the experiment, either individual or social. The second variable represented the experimental condition, either discussion or silence. Finally, we used a time variable, representing the succession of the different measurements.

We used the R software and the package lme4 to build logistic regression models with a logit link function. We analyzed the results of the two problems separately and present the results of a single model that includes the predicted three-way interaction between the three explanatory variables. We conducted only a limited number of planned comparisons in relation to the hypotheses formulated based on the literature and therefore report exact p-values (alpha set at 5%).

Results

As predicted, we found a significant three-way interaction for both problems (Bat and Ball: GLMM, X^2 (df) = 0.6, p = .001; Linda: GLMM, X^2 (df) = 35.2, p < .001; see Figure 1; details of all the GLMM models are provided in supplementary material).

Regarding the Bat and Ball problem, during the Individual phase, we found no difference between conditions in either intercept (Wald test, $\beta_{[silence]} - \beta_{[discuss]}$ = 0.15, SE = 0.39, Z = 0.37, p = .71) or slope (Wald test, $\beta_{[silence]} - \beta_{[discuss]}$ = 0.058, SE = 0.13, Z = 0.44, p = .66). In both conditions the odds of success significantly increased over time (in the Silence condition Wald test, $\beta_{[time]}$ = 0.34, SE = 0.09, Z = 3.74, p < .001; in the Discussion condition Wald test, $\beta_{[time]}$ = 0.40, SE = 0.10, Z = 4.14, p < .001). Interestingly, in the Silence condition there was a sign of a reduction in the increase in success over time during the Social phase compared to the individual phase (Wald test, $\beta_{[Phase=social*time]}$ = -0.19, SE = 0.09, Z = -2.05, p = .04). In this condition, the odds of success increased over time by an estimated 16% per time interval (Wald test, $\beta_{[time]}$ = 0.15, SE = 0.09, Z = 1.75, p = .08).

By contrast, during the Social phase of the Discussion condition, the odds of success increased over time by an estimated 145% per time interval (Wald test, $\beta_{[time]}$ = 0.90, SE = 0.18, Z = 5.08, p < .001). The increase was significantly superior to the increase measured in the Individual phase of the Discussion condition (Wald test, $\beta_{[Phase=individual*time]}$ = -0.50, SE = 0.19, Z = -2.62, p = .009), as well

as the Social phase of the Silence condition (Wald test, β*[Condition=silence*time]* = -0.75, *SE* = 0.20, *Z* = -3.79, $p < .001$).

We found a similar but even clearer pattern of results for the Paul and Linda problem. During the Individual phase, we found no difference between conditions in either intercept (Wald test, β*[silence]*-β*[discuss]* = 0.04, *SE* = 0.36, *Z* = 0.12, $p = .90$) or slope (Wald test, β*[Condition=discuss*time]* = -0.20, *SE* = 0.19, *Z* = -1.06, $p = .29$). In both conditions the odds of success decreased over time (in the Silence condition Wald test, β*[time]* = -0.17, *SE* = 0.13, *Z* = -1.37, $p = .17$; in the Discussion condition Wald test, β*[time]* = -0.37, *SE* = 0.14, *Z* = -2.71, $p = .007$). This decrease likely reflected the fact that participants were pressed to give an answer to a multiple choice question after 20 seconds and therefore that they were responding almost at chance (chance level was 33%, average success at the first response was 27%).

During the Social phase there was no sign of improvement over time in the Silence condition (Wald test, β*[time]* = 0.02, *SE* = 0.13, *Z* = 0.20, $p = .84$). By contrast, we found a sharp increase in success in the Discussion condition, in which the odds of success increased over time by an estimated 151% during the Social phase (Wald test, β*[time]* = 0.92, *SE* = 0.15, *Z* = 6.03, $p < .001$), a significant difference from the Individual phase (Wald test, β*[Phase=individual*time]* = -1.29, *SE* = 0.15, *Z* = -8.67, $p < .001$) and from the Social phase of the Silence condition (Wald test, β*[Condition=silence*time]* = -0.89, *SE* = 0.20, Z = -4.49, $p < .001$).

Remarkably, in one of the discussion group none of the participants had the correct answer at the end of the individual phase. As predicted, the group remained stuck on the incorrect answer during the discussion phase, leading to an average score of 0 at the end of the experiment. By removing this group from analysis we can get a better estimate of the effect of having at least one participant with the correct answer in a discussion group. Without the unanimously wrong group, the estimated increase in success with time rises from 151% to 186% during the Discussion phase.

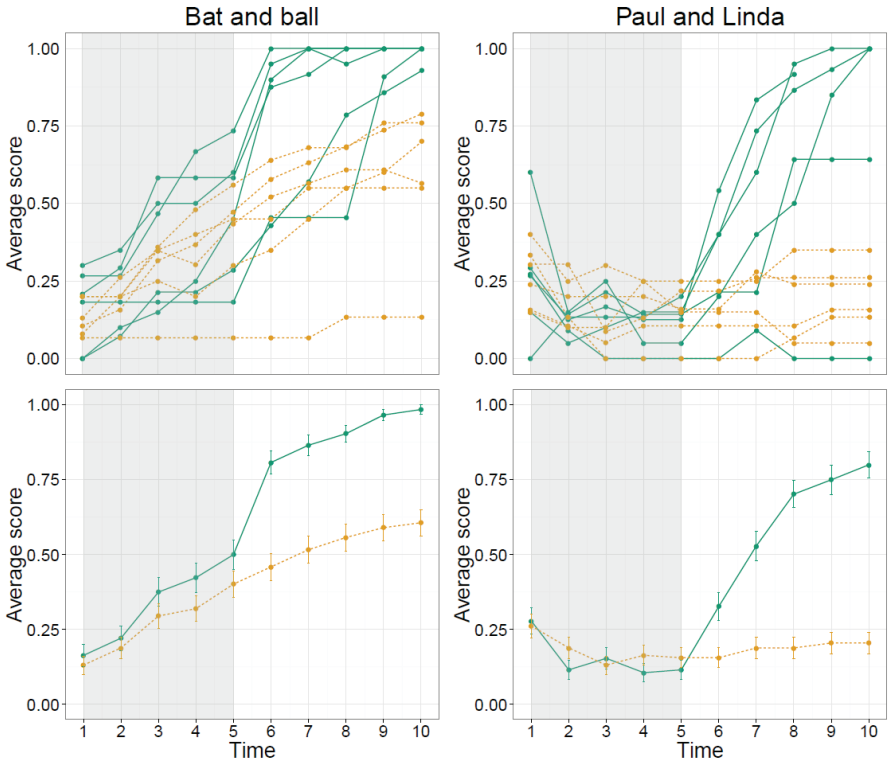


Figure 1: Evolution of the average score during the Individual (shadowed area) and Group phases (clear area) in the Silence (yellow) and Discussion (green) conditions. Top: average for every group. Bottom: average over all groups. Error bars represent standard error.

Discussion

Discussion among participants enabled the spread of the counter-intuitive but correct answer for both problems. As long as some group members had understood the correct answer, they were able to convince their neighbors, who could convince their neighbors in turn until the whole group had accepted the correct answer. The correct answer spread even when it was initially defended by a small minority of participants (2 out of 14 in one group, 3 out of 24 in another), and in spite of the fact that most of the participants who initially defended the wrong answer did so very confidently (the modal confidence at the end of the individual was 10—the maximum—for both problems [Paul and Linda, *Mean* = 8.8; Bat and Ball, *Mean* = 8.0]).

The benefits of discussion are especially striking when compared with the lack of benefits from the mere knowledge of others' answers. Being able to see the other participants' answers yielded no improvement in performance: in neither problem was the rate of increase in correct answers higher in the Social phase than in the Individual phase. These results show that discussion can spread the correct answer but social relationships as well as arguments play an important role in an open setting. The following experiments were designed to further test the hypothesis that argumentation is necessary and sufficient to spread counter-intuitive beliefs.

Study 2: Effect of repeated transmission on the quality of arguments

Study 2a: robustness of a single argument to repeated transmission

While face-to-face argumentation can efficiently spread counter-intuitive beliefs, it remains limited in its speed, and potentially in its scope compared with mass media for instance. People acquire and transmit information through formats that can reach a large population, but that do not allow the extensive back and forth of a face-to-face discussion. Can these formats also enable the wide spread of counter-intuitive beliefs? For this to be possible, two conditions must be met. The first is that people should be convinced by a single argument instead of a discussion. Previous experiments have shown that a substantial number of participants can be convinced to accept a counter-intuitive belief in these conditions (Trouche et al., 2014).

The second condition is that people should be able to convince others in turn, and that the people they convince should be able to convince others, and so forth. This has never been demonstrated so far. To test this, we rely on a variation of the method of transmission chain (e.g. Bangerter, 2000; Bartlett, 1932; Mesoudi & Whiten, 2008) in which a first generation is given an input that it must recall, the recalled input is used as input for another generation, and the process is iterated for several generations. Bartlett used this technique with an argument, but he did not measure the evolution of the persuasiveness of the argument as the number of transmission events increases. In this experiment, the first generation of participants is given a reasoning problem and asked to provide an answer and an argument defending this answer. Participants of the second generation are asked to solve the same problem, are provided the answer and argument for the correct answer from a participant from the first generation. They are then given the chance to change their mind, and asked to provide an argument for their final answer. This process is iterated eight times, for a total of eight generations.

Participants

453 participants were recruited through Amazon Mechanical Turk (281females, M_{Age} = 31.3, *SD* = 10.8). They were paid $0.5.

Materials

The participants all completed the Paul and Linda problem described above.

Procedure

The first generation was composed of 30 participants who were given the problem, asked to provide an answer and an argument for their answer.[5] Their justifications were coded according to the following scheme:

0 = Incomplete argument for the good answer (e.g. "Not matter Linda's marital status this would be true").

[5] All participants filled in demographic information at the end of the experiment. They were also asked about their confidence in their answer but these results are not discussed further here.

1 = Complete argument for the correct answer (e.g. "If Linda is married, she is looking at Patrick, who is not married. If Linda is not married, then Paul (who is married) is looking at her. Thus, in either case, someone who is married is looking at someone who is not married. The answer is Yes.").

2 = Any incorrect argument for the correct answer (e.g. "I would convince someone that the answer is yes. Paul is already married and Patrick is not married. Linda must also not be married, since she is looking at Patrick. Maybe she desires him. Since that is the case, Paul is looking at someone who is not married.").

3 = Standard argument for the common wrong answer (e.g. "It doesn't say weather Linda is married therefore you cant say for sure.").

4 = Other arguments for the common wrong answer, and arguments for the other wrong answer (e.g. "We have no idea who people are actually looking at.")

Participants from generation 2 had to provide an initial answer to the same problem, and were then given an answer and argument presented—truthfully—as coming from a participant in a previous experiment. Four complete arguments for the correct answer (code 1) were selected at random from generation 1 and each participant from generation 2 was randomly assigned one of this four arguments. Participants then had to give a final answer for which they had to give an argument. From each group we randomly drew an argument that fulfilled the following two conditions: (a) it was a complete argument for the correct answer (code 1); (b) it was given by a participant who had initially provided the common wrong answer ('We can't tell'). Participants from generation 3 were then randomly assigned one of these four arguments (see Figure 2 for an example of transmission for one group). The process was iterated until generation 8 was reached.

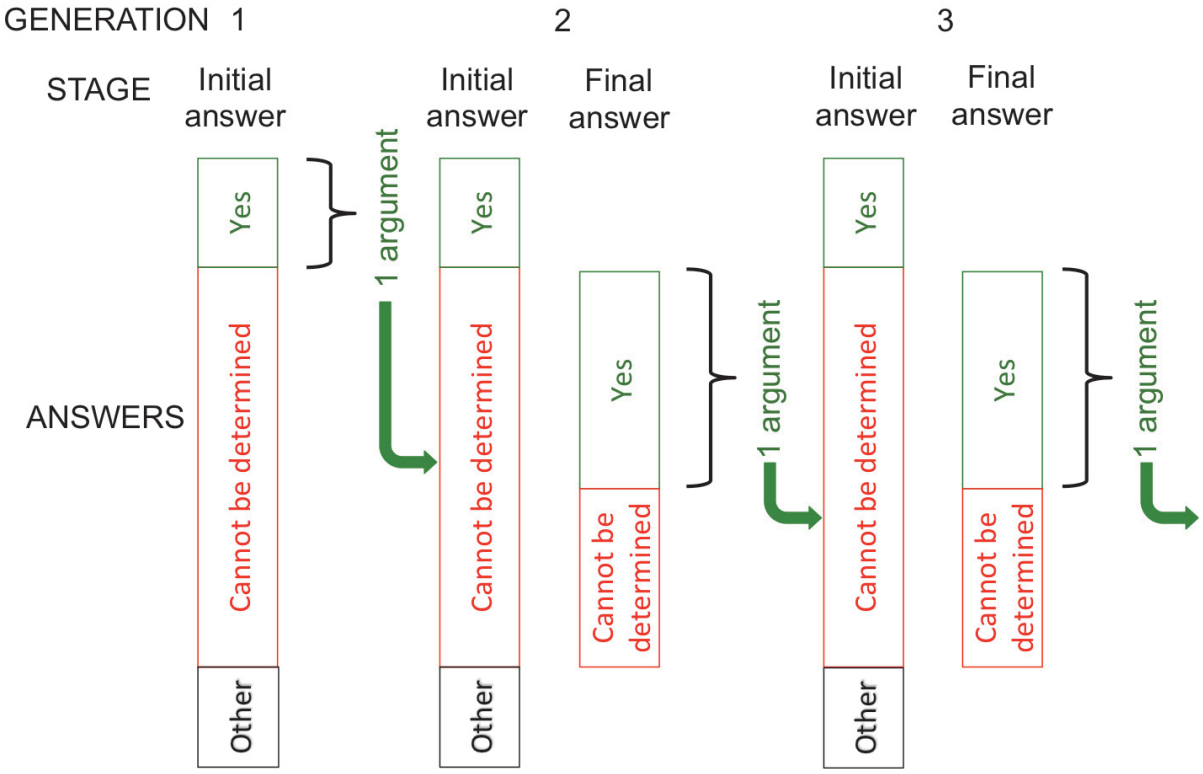


Figure 2: Method for the transmission of arguments in Study 2.

Results

To analyze the results, we rely on three main measures. The first is argument persuasion effectiveness: the proportion of participants who had started with the intuitive wrong answer ('We can't tell') and changed their mind for the correct answer. This might include participants who were then unable to produce a complete and correct justification for the correct answer. The second measure is the transmission effectiveness of the arguments for the correct answer. Argument transmission effectiveness adds to argument persuasion effectiveness the requirement that participants be able to produce a correct and complete argument for the correct answer (code 1). Finally, we measured the difference between these two measures, that is, the proportion of participants who, having been persuaded to change their mind from the wrong to the correct answer, were then able to provide a correct and complete argument. From a cultural evolution perspective these measures play different roles. A very persuasive argument changes the proportion of individuals adopting the correct answer. If the argument changes the mind of enough individuals, it could then have a further impact through conformity. However, the argument itself might not be transmitted, and this might hamper or even preclude further impact if the correct answer needs to be supported by the correct argument to spread. By contrast, an argument with high transmission effectiveness will both persuade individuals to adopt the correct answer and allow them to transmit the argument adequately. In each study, we only present the results of the most relevant measure(s).

Study 2a is meant to test the robustness of argument transmission. The most relevant measure is thus argument transmission effectiveness, since it represents the capacity of an argument to be clearly

understood and to lead to the production of a similar argument. Results show no change in argument transmission effectiveness over time (GLMM Wald test, $\beta[time]$ = -0.02, SE = 0.08, Z = -0.25, p = .80; see Figure 3; details of all the GLMM models are provided in supplementary material). This suggests that when participants were convinced by a correct argument, the argument they put forward to defend the correct answer was as convincing as the argument which had convinced them.

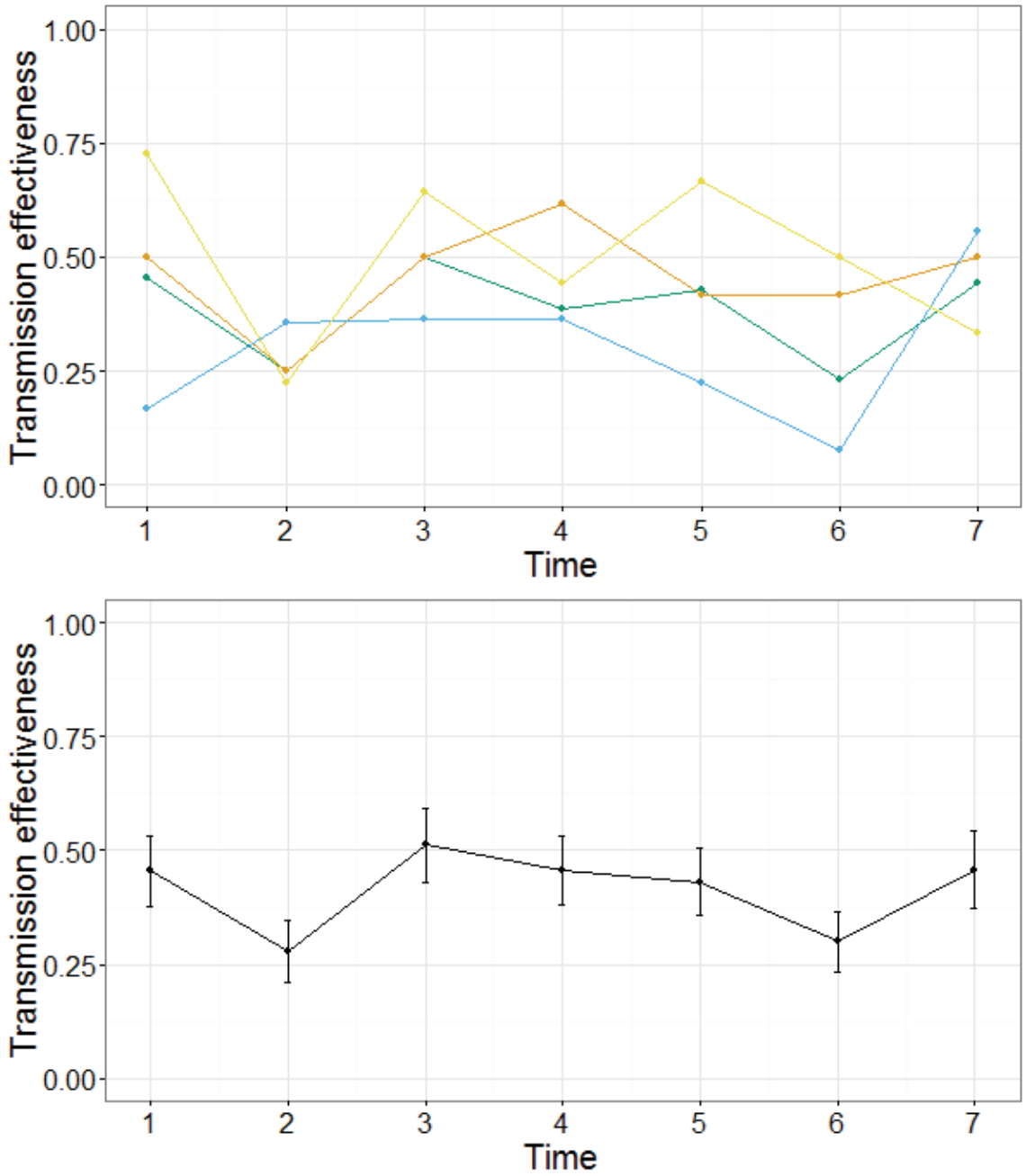


Figure 3. Transmission effectiveness of the arguments from the four chains of Study 2a. Top: by chain. Bottom: on average. Error bars represent standard error.

This outcome could be obtained through two mechanisms. Participants could memorize the argument they have received with high fidelity, or they could reconstruct the argument on the basis of their newly acquired understanding of the problem. A superficial examination of the arguments provided by the participants suggests that they were not simply memorizing the arguments they had received, since their superficial features were often significantly modified, as in the following example:

Argument received: "If Linda married - She is looking at Patrick who is not married – yes. If Linda not married - Paul is looking at Linda – yes"

Argument produced: "There are only two possibilities either Linda is married or she is not married. We know Linda is looking at Patrick and Patrick is not married, so if Linda is married then someone who is married is looking at someone who is not married. What if Lind is not married? Well, Paul is looking at Linda and Paul is married, so if Linda is not married we still have someone who is married, namely Paul, looking at someone who is not married, namely Linda. Either way someone who is married is looking at someone who is not married."

As is apparent in this example, in some cases the arguments produced were significantly more elaborate than the arguments received, which is suggestive of a process of reconstruction. To further demonstrate the importance of reconstruction in argument production, in Study 2b we provided participants with incomplete arguments to see if they would then produce equally incomplete arguments or if they would reconstruct them in a more complete form.

Study 2b: reconstruction of incomplete arguments

Study 2a was conservative in that incomplete arguments for the correct answer were not counted as correct answers. However, these arguments can be used to test how much people reconstruct vs. memorize the arguments that have convinced them. The prediction is that incomplete arguments will be less convincing, since they are harder to understand, but the participants who are convinced will have had to mentally reconstruct the whole reasoning behind the correct answer. It is then possible that when asked to produce an argument in turn, they produce a complete rather than an incomplete argument.

Participants

44 participants were recruited through Amazon Mechanical Turk (24 females, M_{Age} = 35.2, SD = 11.5). They were paid $0.5.

Methods

Three incomplete arguments (code 0) were randomly drawn from the arguments of Study 2a. Each was given to new participants using the procedure used for generations 2 to 8 of Study 2a. As an example, the argument used to illustrate code 0 above was one of the three arguments used.

Results

We first examined argument transmission effectiveness, measured in the same way as in Study 2a. As expected, the participants were less likely to be convinced and to produce a complete argument from an incomplete argument (24% transmission effectiveness) than from a complete argument (45% transmission effectiveness) (albeit non-significantly so: X^2 (1, N = 44) = 3.04, p = .08). However, of the participants who were convinced by the correct but incomplete argument, the majority (9 out of 12) produced not only correct but also complete arguments. The proportion of participants producing such arguments was not significantly lower than with a complete argument (Study 2a, X^2 (1, N = 44) = 0.02, p = .88). Here is for instance the argument produced by one of the participants who had been convinced by the argument used:

Argument received: Not matter Linda's marital status this would be true

Argument produced: The answer is 'Yes someone who is married is looking at someone who is not married'. This statement is true regardless of Linda's marital status because if Linda IS married, then her looking at Patrick (who is NOT married) would satisfy the statement; If Linda is NOT married, then Paul (who IS married) looking at Linda would satisfy the statement.

These results show that participants can reconstruct a complete argument from an incomplete one and suggest that reconstruction is also crucial when participants are provided with a complete argument to start with. This process of reconstruction likely explains the remarkable robustness of the arguments.

Study 3: Effect of source based biases on the diffusion of counter-intuitive beliefs

Studies 1, 2a, and 2b demonstrated the efficacy with which argumentation can spread counter-intuitive beliefs. They also suggested that argumentation could potentially overcome other factors such as conformity: in Study 1, the counter-intuitive but correct answer defended by a minority of participants spread against the intuitive but incorrect answer defended by the large majority. However, this does not show that social factors such as prestige or conformity could not also spread counter-intuitive beliefs. In the following studies, we test whether conformity and prestige can either contribute to the spread of these beliefs on their own, or assist argumentation in spreading these beliefs.

Study 3a: effect of pure conformity

Participants

121participants were recruited through Amazon Mechanical Turk (48 females, M_{Age} = 32.2, SD = 10.0). They were paid $0.5.

Methods

The methods were similar to those of Study 2b except that instead of being provided with the answer and argument of a previous participant, participants were provided with the number of answers of a group of 50 participants described has having previously completed the same problem. In the Majority

Correct condition, 45 of these participants had answered 'Yes' and 5 had answered 'We cannot tell'. In the Majority Incorrect condition, the numbers were reversed. In two more conditions (Majority Correct with Arguments and Majority Incorrect with Arguments), participants were also provided with one argument supporting each of the two answers, ostensibly given by some of the previous participants (fully between-participant 2 x 2 design).

Results

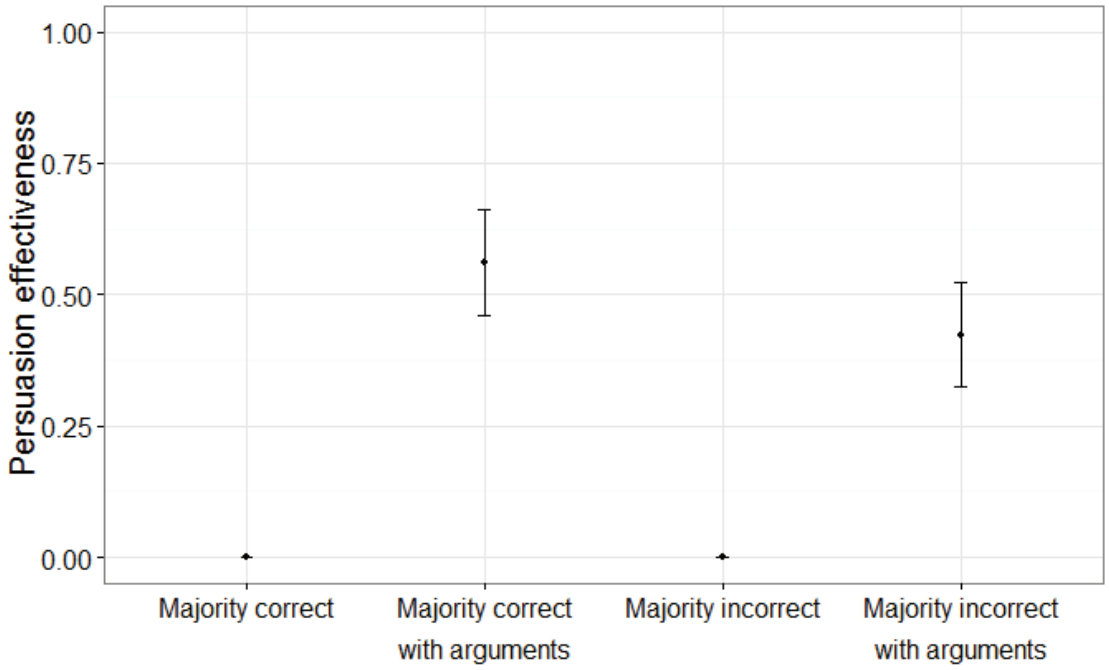


Figure 4: Effect of conformity with and without argument on persuasion effectiveness. Error bars represent standard error.

Compared to Studies 2a and 2b, here we are interested first in the persuasive impact of the information provided by the participants—that is, does it make them adopt the correct answer, irrespective of whether they understand the argument or not (persuasion effectiveness).

When only the majority information was presented, it had no effect on persuasion effectiveness, whether the 'Yes' (correct) answer was described as being held by a majority or a minority of previous participants (X^2 (1, N = XX) = 0.34, p = .56) (see Figure 4). By contrast, introducing the arguments significantly raised the persuasion rates (X^2 (1, N = XX) = 28.40, p < .001). Finally, participants were more likely to be persuaded when the correct argument was that of the majority rather than the minority, but this difference was small and far from significance (X^2 (1, N = XX) = 0.49, p = .49).

Study 3b: effect of exposition to several arguments

The goal of Study 3b was to test the robustness of the conclusion from Study 3a that, to spread the counter-intuitive beliefs under study, conformity plays a negligible role by contrast with argumentation.

Another form of conformity could come from the number of arguments presented: participants might not be sensitive to the fact that many people agree on an answer, but still be sensitive to the fact that many people produce similar arguments. We thus manipulated the proportion of correct arguments participants were exposed to, from one out of five to five out of five. If conformity supported argumentation, we would expect argument persuasion effectiveness to increase with the proportion of correct arguments presented.

Participants

153 participants were recruited through Amazon Mechanical Turk (64 females, M_{Age} = 30.6, SD = 10.5). They were paid $0.5.

Methods

The methods were similar to those of the first generation of Study 2b except that instead of being provided with the answer and argument of a single previous participant, participants were provided with the answer and argument of five previous participants. The number of correct argument varied from one to five, and the arguments were presented in three different orders (5 x 3 between-participant design).

Results

As in Study 3a we analyzed argument persuasion effectiveness. The order in which the arguments were presented had no impact on argument persuasion effectiveness (GLMM, X^2 (2) = 1.17, p = .56). More importantly, neither did the number of correct arguments (GLMM, X^2 (1) = 0.0003, p = .99 (see Figure 5).

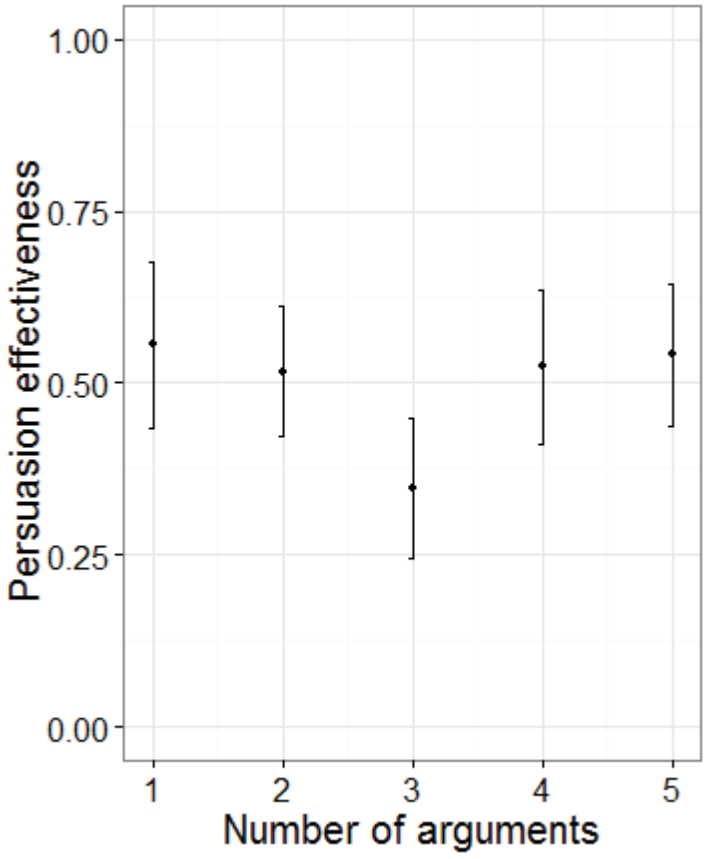


Figure 5: Effect of the number of arguments on argument persuasion effectiveness in Study 3a. Error bars represent standard error.

Study 3c: effect of prestige

In study 3a and 3b we showed that conformity has a very limited effect on the efficiency of arguments and the spread of counter-intuitive answers to our problem. However, conformity is a property of groups: one individual changes her behavior based on a majority of other individuals in the group. By contrast, prestige comes from the social status of a few particular individuals and as a consequence many individuals may change their behavior based on these few individuals (Boyd & Richerson, 1985). As a result, if a prestigious individual offers a counter-intuitive answer and an argument supporting it we might expect participants to accept the answer and the argument, and to transmit them to other participants in order to convince them. The following study therefore tested the possibility that prestige can lead to the spread of counter-intuitive answers and of the arguments supporting them.

Participants

60 participants were recruited through Amazon Mechanical Turk (40% females, M_{Age} = 33.6, *SD* = 11.4). They were paid $0.5.

Methods

The methods were similar to those of Study 2a except that instead of being provided with the answer and argument of a previous participant, participants were explicitly told by the experimenters that they were given the correct answer to the problem (Pure Prestige condition). In the Prestige and Argument condition, the correct answer was accompanied by a correct, complete argument (between-participant design).

Results

As in Study 3a, we first analyzed persuasion effectiveness. In both conditions, this proportion was superior to 0.5 (Figure 6). Although persuasion was higher in the Prestige and Argument condition than in the Pure Prestige condition, this difference did not reach significance (X^2 (1, *N* = XX) = 0.85, p = .36). These results show that a strong enough prestige cue can lead people to accept a counter-intuitive belief.

However, only few of the participants who accepted the correct answer in the Pure Prestige conditions were able to provide a correct and complete argument for their answer (3 out of 11 participants, 27%). The proportion of complete arguments was smaller in the Pure Prestige condition than in the Prestige and Argument condition (7 out of 11 participants, 63%, albeit non-significantly so (X^2 (1, *N* = XX) = 1.65, p = .20) and significantly smaller than when participants were simply given an argument for the correct answer without information on prestige (Study 2a, 76.4%; X^2 (1, *N* = XX) = 10.3, p = .001).

Moreover, none of the participants who had accepted the correct answer on the basis of prestige mentioned the source that influenced them to justify their answer (i.e. "The experimenters told us that

was the right answer"). This suggests that even if prestige can help spread a counter-intuitive belief in the coterie of the prestigious source, it cannot help spread it any further.

To confirm this finding, we carried out two supplementary studies. In the first, we selected at random three of the incorrect arguments for the correct answer provided by participants who had changed their mind on the basis of prestige only. Out of the 35 participants (from the same population as studies 2 and 3) who started out with the intuitive but wrong answer, only one ended up with the correct answer and the correct justification.

Even though the participants had not spontaneously provided any argument from authority, in the second supplementary study we also checked whether they would have been efficient. We created an argument from authority ("The experimenters told us that was the right answer") and gave it to participants (again from the same population) as being the argument given by another participant in support of the 'Yes' answer. Out of the 28 participants (from the same population as studies 2 and 3) who started with the intuitive but wrong answer, four changed their minds, but none produced a correct argument, and only one repeated an argument from authority.

These results thus suggest that the effects of prestige are strictly limited to the participants who are in immediate contact with the prestigious source, and that prestige could not enable the diffusion of the counter-intuitive beliefs under study.

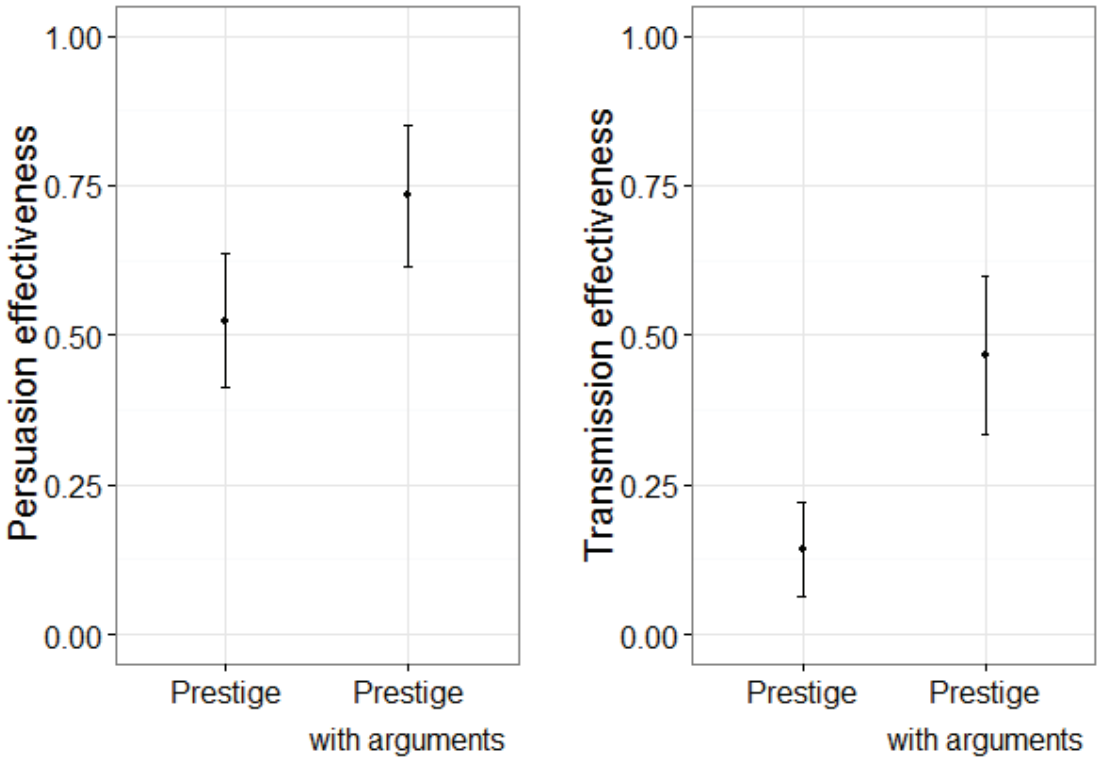


Figure 6: The effect of prestige on the probability that initially wrong participants are persuaded by a correct argument (left) and able to produce a correct and complete justification (right). Error bars represent standard error.

Conclusion

Even though intuitive beliefs spread more easily than fully counter-intuitive beliefs, the latter has still been observed to spread in some cases—scientific theories being the most striking example. We hypothesized that argumentation could play major role in spreading counter-intuitive beliefs and produce an original evolutionary dynamic. To test this hypothesis in controlled conditions, we relied on simple reasoning problems that have an intuitive but wrong solution, and a counter-intuitive but correct solution. In a series of experiments, we demonstrated the power of argumentation to spread the counter-intuitive but correct solution.

Study 1 showed that when participants can discuss the problems together, the correct answer spreads very effectively, even in groups larger than those usually studied. Participants who initially find the correct answer are able to convince the participant they discuss with to accept it. Crucially, the participants who have been convinced can then convince others in turn, until the answer is accepted by the whole group. This is true even when the participants who initially defend the correct answer are a minority, and when they face a confident majority.

Study 2a further demonstrated the robustness of argument transmission. Participants were asked to complete a logical problem, were provided with an argument for the correct answer, could change their mind on this basis, and had to produce an argument for their final answer. The arguments produced were then used as input for another generation of participants, for a total of eight generations. There was no loss of quality in argument effectiveness: the participants who changed their mind produced arguments that were just as convincing as the argument that had convinced them. Study 2b further showed that the robustness of argument transmission was due to the reconstruction of arguments by the participants. Instead of memorizing the argument that had convinced them, the participants reconstructed an argument on the basis of their new understanding of the problem. Participants who managed to understand the problem on the basis of an incomplete argument produced correct and complete arguments.

Studies 3a, 3b, and 3c showed that conformity or prestige do not have the same effects as argumentation on the transmission of counter-intuitive beliefs. When participants were told that a majority of participants had answered in a given way, none changed their mind, unless they were also provided with an argument for the correct answer. Even then, they were not significantly more likely to change their mind than if the argument had been presented without the majority information. Similarly, presenting participants with five arguments for the correct answer from previous participants had no more effect than presenting them with one argument for the correct answer and four arguments for the intuitive but wrong answer.

When we told participants, as experimenters, what the correct answer was, approximately half accepted it on the basis of prestige. However, very few participants were then able to produce a convincing argument for the correct answer, and none used arguments that would allow the effects of

prestige to be transmitted to another generation of participants. Participants told, explicitly by us, what the correct answer was, but also given an argument for the correct answer, were more likely to then produce correct arguments. However, again none of their arguments referenced the initial source of authority, so that any effect of prestige would be lost after one generation.

Taken together, these results show that argumentation can effectively spread counter-intuitive beliefs and that, by contrast, conformity and prestige have limited effect on the spread of such beliefs, at least beyond the immediate vicinity of prestigious sources. Study 1 also suggested that when conformity and argumentation were pit against each other—as the correct answer was defended by a small minority of participants—then conformity did not significantly hinder the spread of counter-intuitive beliefs. This conclusion is reinforced by recent results showing that strong source based cues—such as the benevolence or the competence of the source—do not stop people from accepting strong enough arguments (Trouche, Shao, & Mercier, submitted; see also Castelain, Bernard, Van der Henst, & Mercier, in press).

To the best of our knowledge, the only past experiment to have investigated the behavior of arguments in transmission chains was that of Bartlett (1932). As mentioned above, in this experiment the arguments were nearly entirely lost after a few generations. This might be explained by the fact that the arguments were relatively long, that they had no particular relevance for the participants, and that the participants were asked to memorize them, not to use them in an attempt to convince someone else. By contrast, in our experiments the arguments where short, were relevant to the participants, and the participants had to produce new arguments to convince someone else. From a methodological point of view, this last point in particular bears emphasizing. Contrary to most prior studies of transmission chains in humans, what was tested here (in particular in Study 2a) was not whether a given piece of information was faithfully reproduced from one generation to the next, but whether this piece of information has the same effect on others from one generation to the next. This focuses the study on the most relevant traits of the piece of information in a given context—here, how convincing an argument is.

In order to better understand the evolutionary dynamic at play in the present experiments, we constructed simple models based on the results from Studies 2 and 3. These models are useful to extrapolate and generalize from the data obtained in the experiments and to represent situations that fall outside the experimental context. Following the experiments, the models represent a population of individuals (fixed at 1000 individuals) that attempt to solve a counter-intuitive problem on their own, and are then exposed to one or more arguments randomly coming from participants from the previous generation. At each generation, the entire pool of individuals changes. The rate of correct arguments the individuals can find on their own (during the individual phase of the experiments) is fixed at 10% to approximate the low rate of correct arguments typical of counter-intuitive problems. We study the effects of two variables. First, the transmission effectiveness (TE) of the arguments: the probability that when exposed to a correct argument an initially wrong participant accepts the correct answer and produces a correct argument. Second, the number of arguments (N) an individual is exposed to. We assume that if

there is at least one correct argument among the arguments participants receive, they are convinced and able to transmit the argument with a certain probability (i.e. TE). We also assume that individuals exposed to wrong arguments are never convinced to abandon the correct argument (which was the case in our experiments). The main outcome of interest is the proportion of individuals with the correct argument at equilibrium (when this proportion is stable).

Figure 7 summarizes the findings. When individuals are exposed to a single argument (N = 1), only very high transmission effectiveness (TE ≥ 90%) generate equilibria above 50% (Fig. 7, A). This means that for the correct answer to spread significantly along chains of single participants, transmission must be very efficient. By contrast, with a transmission effectiveness of 50%, an equilibrium above 50% is reached when individuals are exposed to three or more arguments (Fig. 7, B). This illustrates the importance of redundancy in cultural transmission (see, Acerbi & Tennie, 2016; Eriksson & Coultas, 2012; Morin, 2015; Muthukrishna, Shulman, Vasilescu, & Henrich, 2014).

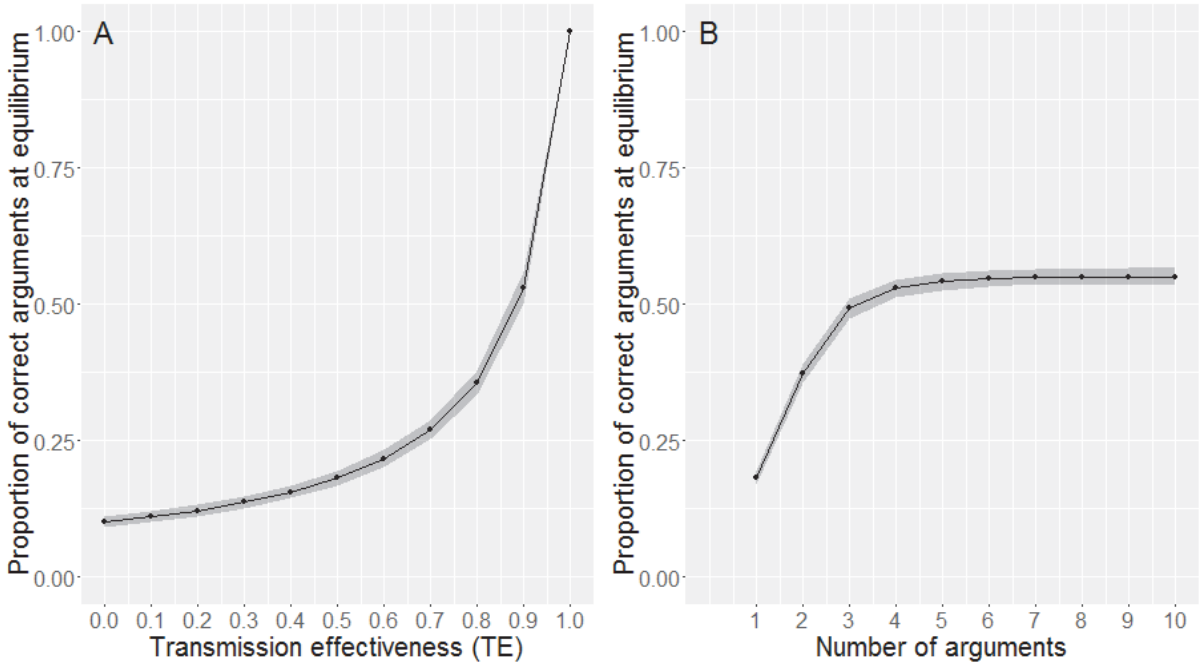


Figure 7: Effect of transmission effectiveness (TE) and the number of arguments (N) on the spread of counter-intuitive beliefs (proportion of participants giving a correct argument at equilibrium). A: When the number of arguments is fixed at one (N = 1), transmission effectiveness must be high to spread the arguments. B: The number of models (N) has an important impact on the diffusion of counter-intuitive beliefs when transmission effectiveness is limited (TE = 50%). Dots represent mean values of 100 simulations, grey shading the standard deviation around the mean. Other parameters of the simulation are the number of individuals (1000) and the probability to find a correct argument during the initial phase (10%).

What are the cognitive mechanisms that enable argumentation to spread counter-intuitive beliefs? One possibility is that argumentation makes counter-intuitive beliefs temporarily intuitive. For instance, with the problems used here participants exposed to the correct argument for the correct answer often

immediately and intuitively grasp why the argument is correct, and thus why the answer it supports is correct (see, e.g., Mercier & Sperber, 2011). That is, when participants are confronted with the argument, the correct answer becomes the conclusion of a succession of intuitive inferential steps. Note that this does not mean that the correct answer remains intuitive. On the contrary, the wrong answer often keeps exerting an intuitive pull (see, e.g., Sloman, 1996), and it is likely that participants have to reconstruct the reason for why the correct answer is correct every time they want to convince themselves or someone else of its correctness.

A limitation of our experimental approach is that we study beliefs that are much simpler than culturally significant beliefs such as counter-intuitive scientific or theological theories. We could hardly let people talk with each other and wait for them to discover on their own and spread the theory of evolution by natural selection or Bayes' theorem. As a result, our participants have relatively little commitment to their intuitive but wrong beliefs, which might help explain why argumentation so efficiently spreads the correct but counter-intuitive answer. Indeed, it has been suggested that scientists' commitments to their beliefs could be so strong as to make them deeply reluctant to endorse revolutionary theories (e.g. Kuhn, 1962).

Yet, in spite of these potential difficulties, well-supported scientific theories spread very quickly, even when they are revolutionary. This is particularly true in mathematics. For instance, Gödel's incompleteness theorem was promptly accepted by the mathematical community, even though it disproved beliefs on which eminent members of the community—such as Hilbert or Russell—had wagered their careers (Mancosu, 1999). In the sciences, revolutionary theories also spread very quickly, sometimes only taking a few years to go from eccentric hypotheses to textbook examples (e.g. Oreskes, 1988). Indeed, it has been argued that revolutionary scientific theories spread in the relevant community about as quickly as is warranted by the evidence garnered in their support (Kitcher, 1993; Mercier & Heintz, 2014; Wootton, 2015). Crucially, as noted in the introduction, this is true even if the defenders of the new theories have no special status and are, by definition, a small minority. It is thus possible that our simple experiments capture an important dimension of the spread of counter-intuitive beliefs outside the laboratory.

Bibliography

Acerbi, A., & Tennie, C. (2016). The role of redundant information in cultural transmission and cultural stabilization. *Journal of Comparative Psychology*, *130*(1), 62.

Bangerter, A. (2000). Transformation between scientific and social representations of conception: The method of serial reproduction. *British Journal of Social Psychology*, *39*(4), 521–535.

Bartlett, S. F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago: Chicago University Press.

Boyer, P. (2001). *Religion Explained*. London: Heinemann.

Castelain, T., Bernard, S., Van der Henst, J.-B., & Mercier, H. (in press). The influence of power and reason on young Maya children's endorsement of testimony. *Developmental Science*.

Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1797), 20141541.

Claidière, N., & Sperber, D. (2010). Imitation explains the propagation, not the stability of animal culture. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1681), 651–659.

Eriksson, K., & Coultas, J. C. (2012). The advantage of multiple cultural parents in the cultural transmission of stories. *Evolution and Human Behavior*, *33*(4), 251–259.

Feher, O., Wang, H., Saar, S., Mitra, P. P., & Tchernichovski, O. (2009). De novo establishment of wild-type song culture in the zebra finch. *Nature*, *459*(7246), 564–568.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1509), 3503–3514.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. New York: Oxford University Press.

Kuhn, T. (1962). *The Structure of Scientific Revolutions* (50th anniversary edition). Chicago: Chicago University Press.

Laland, K. N., & Williams, K. (1997). Shoaling generates social learning of foraging information in guppies. *Animal Behaviour*, *53*(6), 1161–1169.

Laughlin, P. R. (2011). *Group problem solving*. Princeton: Princeton University Press.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*, 177–189.

Mancosu, P. (1999). Between Vienna and Berlin: The immediate reception of Godel's incompleteness theorems. *History and Philosophy of Logic*, *20*(1), 33–45.

Maxwell, R. S. (1936). Remembering in different social groups. *British Journal of Psychology. General Section*, *27*(1), 30–40.

Mercier, H., & Heintz, C. (2014). Scientists' argumentative reasoning. *Topoi*, *33*(2), 513–524.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Mesoudi, A., & Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of Cognition and Culture*, *4*(1), 1–24.

Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3489–3501.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to Tango. *Personality and Social Psychology Bulletin*, *37*(10), 1325–1338.

Miton, H., Claidière, N., & Mercier, H. (2015). Universal cognitive mechanisms explain the cultural success of bloodletting. *Evolution and Human Behavior*, *36*(4), 303–312.

Morin, O. (2013). How portraits turned their eyes upon us: Visual preferences and demographic change in cultural evolution. *Evolution and Human Behavior*, *34*(3), 222–229.

Morin, O. (2015). *How Traditions Live and Die*. New York: Oxford University Press.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning*, *4*(3), 231–248.

Muthukrishna, M., Shulman, B. W., Vasilescu, V., & Henrich, J. (2014). Sociality influences cultural complexity. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1774), 20132511.

Northway, M. L. (1936). The influence of age and social group on children's remembering. *British Journal of Psychology. General Section*, *27*(1), 11–29.

Oreskes, N. (1988). The rejection of continental drift. *Historical Studies in the Physical and Biological Sciences*, *18*(2), 311–348.

Rahwan, I., Krasnoshtan, D., Shariff, A., & Bonnefon, J.-F. (2014). Analytical reasoning task reveals limits of social learning in networks. *Journal of The Royal Society Interface*, *11*(93), 20131211.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone*. Chicago: University of Chicago Press.

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, *12*(1), 73–89.

Scott-Phillips, T. (in press). A simple (experimental) demonstration that cultural evolution is not replicative, but reconstructive-and an explanation of why this difference matters. *Journal of Cognition and Culture*. Retrieved from http://dro.dur.ac.uk/16582/

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Sperber, D. (1996). Why are perfect animals, hybrids, and monsters food for symbolic thought? *Method & Theory in the Study of Religion*, *8*(2), 143–169.

Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *94*(1), 197.

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.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

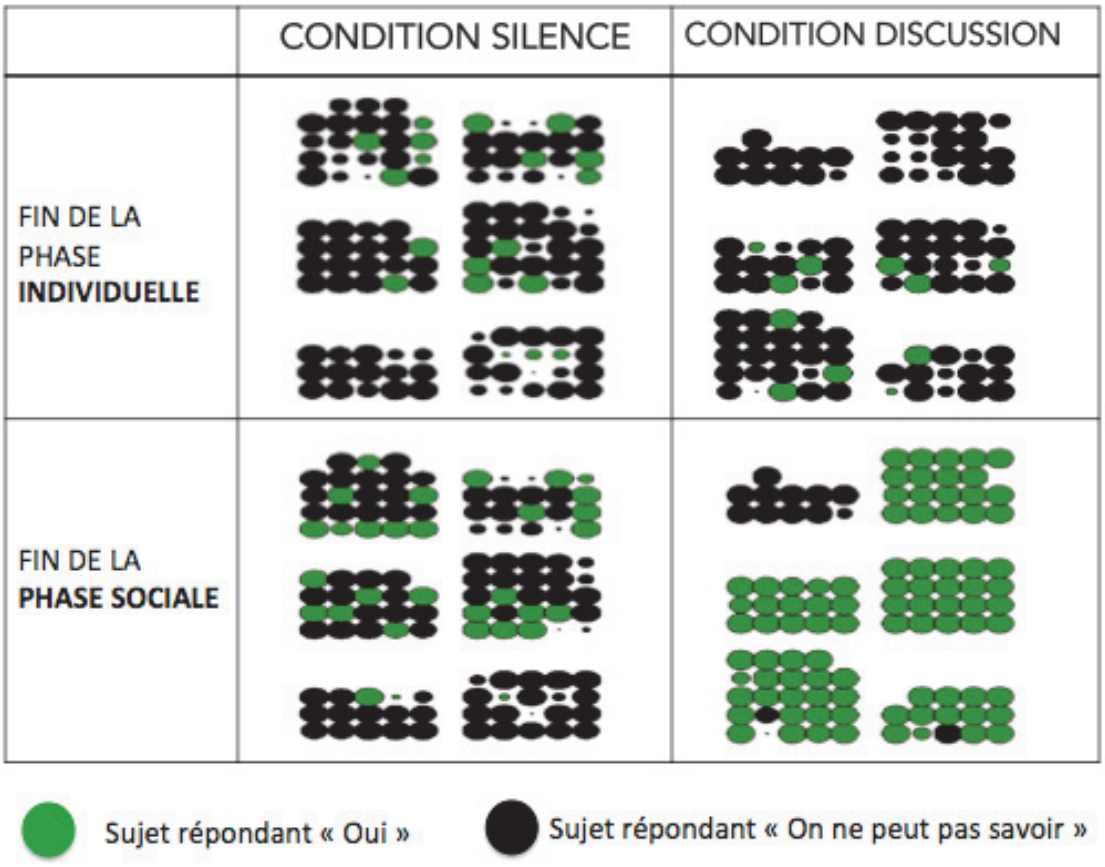
Trouche, E., Shao, J., & Mercier, H. (submitted). How is argument evaluation biased?

Wootton, D. (2015). *The Invention of Science: A New History of the Scientific Revolution*. London: Harper.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York: Springer.

Les résultats de la première étude de l'article précédent montrent très clairement que l'échange d'arguments est essentiel pour la diffusion de la bonne réponse pour ces problèmes. Même dans le cas du bonbon et de la baguette, seuls environ 63% des sujets sont capables de fournir une justification acceptable pour la bonne réponse à la fin de l'expérience, dans la condition silence. Seul 25% dans le cas de Paul et Linda, ce qui contraste avec les résultats de la condition discussion, où quasiment tous les sujets sont capables de fournir une justification acceptable pour la bonne réponse à la fin de l'expérience.

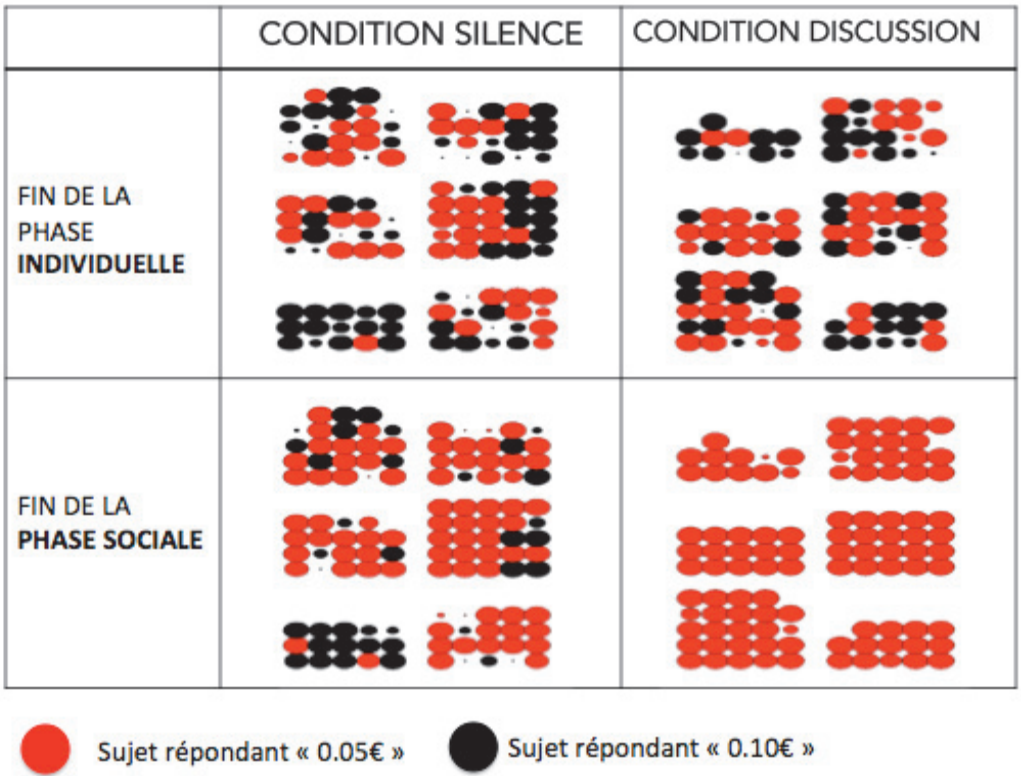
Présentons, en complément de l'article, quelques images réalisées par Nicolas Claidière représentant chaque sujet par un point, la taille des points symbolisant le sentiment de confiance des sujets. Nous avons déjà présenté l'évolution du taux de bonne réponse et de la confiance au cours des quatre minutes et vingt secondes de réflexion personnelle. Voici maintenant à quoi ressemblent les distributions spatiales des bonnes réponses (en vert) et du sentiment de confiance des sujets pour le problème de Paul et Linda, a la fin de la phase individuelle et la fin de la phase sociale dans les deux conditions. Rappelons que chaque groupe de sujet était séparé au moment de l'expérience.



Arrêtons-nous sur le groupe dans la condition discussion dont tous les membres finissent sur-confiants en la mauvaise réponse. Loin d'être une anomalie, ce cas est parfaitement prédit par la théorie argumentative. En effet, le fait d'avoir tous les biais « vers son côté » des membres d'un groupe produisant des arguments dans la même direction et les évaluant superficiellement (car cela conforte leur intuitions) décuple le phénomène de sur-confiance et peut avoir des effets désastreux. C'est ce qui correspond au phénomène de polarisation classique de groupe observé dans la littérature de la psychologie sociale (Wayne et al 2015).

Quand aux deux ou trois sujets qui à la fin de la phase de discussion répondent toujours « on ne peut pas savoir » malgré l'échange d'argument, ce sont bien là, en revanche, des anomalies. Ces sujets ont préféré réinterpréter le problème de manière à ce que leur réponse « On ne peut pas savoir » soit toujours valide malgré l'argument pour la bonne réponse, en défendant l'idée, par exemple, que Linda pourrait être un chien ou un poisson rouge. Dans ce cas là en effet « marié » et « non-marié » ne sont pas les deux seuls cas possible pour le statut de Linda, et l'argument pour la bonne réponse ne tient plus.

Observons à présent les mêmes données dans le cas du bonbon et de la baguette. Les mauvaises réponses, ici « 0.10€ » donc, sont toujours en noir, mais les bonnes réponses sont cette fois en rouge.



Remarquons, dans le cas du bonbon et de la baguette, le groupe en bas à gauche dans la condition silence. A la fin de la phase individuelle, ce groupe de sujet ne contient qu'une seule bonne réponse et seulement deux à la fin de la phase sociale. Cela peut laisser penser que, sans argument, les sujets ne doutent de leur réponse incorrecte qu'à partir du moment où ils sont entourés par plusieurs réponses différentes de la leur. Même dans ces cas là, pendant au moins 3 minutes (le temps minimum après lequel si aucun changement ne s'opère chez les sujets, l'expérience s'arrête), certains sujets se retrouvent majoritairement entourés de réponse correcte « 0.05€ » mais peu d'entre eux semblent douter de leur réponse. Il est intéressant de remarquer toute fois que pour la majorité de ces sujets, ils ont également au moins une réponse « 0.10€ » parmi leur voisin. Ceci peut

laisser penser qu'une sorte de « pression sociale » est en jeu ici et que le fait de ne pas être seul face à celle-ci permet d'y résister, dans ce cas au dépend de leur performances. Pour la plupart des sujets dans l'erreur qui sont isolés, entourés de bonnes réponses, si la majorité change d'avis, certains indiquent simplement un sentiment de confiance plus bas mais ne changent pas d'avis. Sans doute car ils ne comprennent pas les raisons de la réponse 0.05€. On peut penser que leur biais « vers son côté » ajouté à l'évaluation superficielle de leurs arguments sont des biais trop forts qui leur font imaginer des raisons pour lesquelles les autres pourraient avoir tort, plutôt que des raisons pour lesquelles ils auraient tort.

En conclusion de ce chapitre, l'échange d'arguments peut être un moyen puissant d'annuler les effets néfastes du raisonnement individuel à condition que la bonne réponse soit présente chez au moins un des membres du groupe et que celui-ci puisse échanger ses arguments avec les autres. Nous allons, dans le chapitre suivant, proposer une discussion générale à partir des différences entre les deux tâches observées dans ce chapitre. Après une proposition spéculative sur le rôle et le fonctionnement du sentiment de confiance dans les processus de raisonnement, nous passerons au dernier chapitre concluant cette thèse.

Chapitre 7 – Discussion

Existe-t-il des « tâches de raisonnement » ? De notre point de vue : non. Des sujets qui justifieraient un choix perceptuel raisonneraient plus que des sujets qui répondraient « 0.10€ » en quelques secondes et même plus que pour réaliser l'inférence réflexive qu' « on ne peut pas savoir si quelqu'un de marié regarde quelqu'un de non marié » en à peine plus de temps et d'efforts.

Commençons par résumer les différences que nous avons observées entre les tâches de Paul et Linda et du bonbon et de la baguette. Premièrement, au vu du sentiment de confiance des sujets donnant la mauvaise réponse intuitive, il semble que ces sujets ressentent effectivement un sentiment de doute dans le cas du bonbon et de la baguette, cela semble, en revanche, ne pas être le cas dans le cas de Paul et Linda. Deuxièmement, le raisonnement solitaire a plus de chance de corriger les réponses intuitives dans le premier problème que dans le deuxième. La causalité entre ces deux observations n'est cependant pas claire. On pourrait imaginer que c'est ce sentiment de doute qui mène certains sujets à corriger leur erreur intuitive, ou bien simplement que ce sentiment de doute venant d'une erreur de lecture, ce type d'erreur a plus de chance d'être reconnu lors de la justification de la réponse intuitive que lorsque l'erreur est moins facile à trouver, comme dans le cas du non déploiement d'un raisonnement disjonctif qui mène à l'erreur. Les problèmes qui ont des traits de surface suffisamment saillants pour induire un sentiment interne de doute se trouvent également être des problèmes pour lesquels la probabilité de se rendre compte de son erreur en produisant des raisons pour sa mauvaise intuition est plus élevée.

Lorsque ce sentiment de doute est accru par des informations sur ce qu'ont répondu d'autres personnes, comme dans les conditions silence de nos expériences de diffusion, cela augmente encore la probabilité de détection d'erreur dans le cas du bonbon et la baguette. Dans le cas de problème comme Paul et Linda, en revanche, très peu de sujets remettent en question leur réponse malgré le fait que leur voisin réponde « oui ». Il reste cependant difficile de tirer des conclusions fortes à partir de cette comparaison. D'une part, le taux de bonnes réponses est, au démarrage de la phase sociale, bien plus élevé dans le cas du bonbon et de la baguette, ce qui augmente les chances pour les sujets dans l'erreur d'avoir des voisins les faisant douter. De plus, le problème de Paul et Linda est un problème forcément à choix multiples (Oui, Non, On ne peut pas savoir) ce qui rend la réponse différente d'un voisin moins surprenante que dans le cas du bonbon et

de la baguette, sans choix multiples, où la réponse « 0.05€ » n'avait, pour les sujets dans l'erreur, certainement jamais été envisagée.

Retenons donc simplement l'idée suivante : on peut distinguer (ou classer) les tâches incongruentes qui génèrent un sentiment de doute et d'autres où cela ne semble pas être le cas. Nous parlerons dans la suite de tâches conflictuelles, qui tendent à générer un conflit cognitif, même léger, chez les sujets, opposées aux tâches non conflictuelles, qui semblent générer pas ou peu de conflit cognitif chez les sujets.

Remarquons premièrement que les tâches utilisées par De Neys et collaborateurs et discutées par De Neys et Bonnefon (2013) sont toutes des tâches conflictuelles, présentant des éléments suffisamment saillants dans l'énoncé pour générer ce conflit. De plus, pour revenir aux différences individuelles et aux explications en terme d'échec de détection de conflit ou d'échec d'inhibition : de notre point de vue, le sentiment de conflit pousse les gens à continuer la production et l'évaluation de raisons pour leurs intuitions, ce qui augmente les chances qu'ils détectent leur erreur, augmentant par exemple dans le bonbon et la baguette le nombre de lectures de l'énoncé. Dans le cas de Paul & Linda, le conflit cognitif n'est généré que par des informations extérieures (comme l'observation de voisins ayant répondu autre chose). Dans ce cas, on peut imaginer que le sentiment de conflit pousse également les gens à produire et évaluer des réponses pour leur réponse et augmente le nombre de lectures de l'énoncé, mais cela n'augmente que très peu la probabilité de détecter leur erreur pour les problèmes où l'erreur est plus profonde, plus cachée par l'énoncé.

Par ailleurs, faisons une remarque sur ce qui peut créer un conflit cognitif dans notre expérience de diffusion avec le bonbon et la baguette. Au-delà du sentiment interne mesuré par les études de De Neys, que « un processus ne s'est pas déroulé de manière fluide » et de l'observation que nos voisins ont une réponse différente, on peut penser qu'à cela s'ajoute d'autres éléments de la situation générant un sentiment de conflit. En particulier, le fait que le problème, pour les sujets qui répondent 0.10€, peut apparaître comme extrêmement simple, ce qui peut sembler incohérent avec le fait que l'expérimentateur nous donne 4 minutes pour le résoudre et même parfois 20 de plus avec la phase sociale. Par ailleurs, les sujets étant dans la même salle, même lors de la résolution individuelle, il est possible qu'ils détectent que certains de leurs camarades sont en train de changer d'avis.

En conclusion, notre explication des échecs du raisonnement en solitaire pour une tâche comme celle de Paul & Linda se trouve plus dans la non-détection de conflit. Pour le cas de tâche comme le bonbon et la baguette, elle se trouverait non pas dans un échec d'inhibition mais dans la façon de résoudre ce conflit avec le raisonnement, comme mécanisme produisant et évaluant des

raisons. Nous pouvons spéculer, en effet, que la différence entre les sujets qui finissent par corriger leur réponse dans la condition silence et ceux qui la conservent, se trouve dans les raisons qu'ils invoquent pour résoudre le conflit entre leur réponse et celle de certains de leur voisin. Nous pouvons imaginer par exemple, qu'un sujet qui se considère comme bon en mathématiques tende plus facilement à se dire que les autres se trompent qu'un sujet qui s'estime moins bon que la moyenne des gens. Si le sujet est effectivement meilleur en mathématiques, il a de bonnes chances, soit de faire partie des sujets qui ont la bonne réponse, soit de détecter son erreur (par exemple en posant l'équation). En revanche, si ce sujet surévalue quelque peu ses capacités, il risque de résoudre le conflit entre sa réponse et celle des autres par une erreur des autres et non la sienne. Quant aux sujets qui se sentiraient moins bons que les autres pour ce genre de problème, ils ont, eux, toutes les chances de prendre sérieusement en compte les réponses différentes de leur voisin. A moins qu'ils aient vraiment des difficultés sérieuses en lecture ou en mathématiques, ils auraient paradoxalement plus de chance de corriger leurs erreurs que des sujets objectivement meilleur en mathématiques mais globalement trop sûr d'eux. Ces réflexions ne sont bien entendu que des spéculations qui mériteraient des investigations expérimentales adressant directement ces questions de profils de confiance.

Quoiqu'il en soit, l'explication des erreurs des sujets dans des tâches générant un conflit viendrait non pas d'un échec d'inhibition mais dans la façon dont les sujets résolvent leur sentiment conflit, en particulier dans les raisons utilisables par ses sujets. En revanche, comme l'illustre le problème de Paul & Linda l'illustre, pour les tâches qui ne généreraient pas ou très peu de conflit – synonyme que l'erreur est moins détectable -, elles ont très peu de chance de mener le raisonnement à corriger nos intuitions initiales. D'une part, car moins de conflits nous pousse moins à continuer la production et évaluation d'argument pour notre réponse, d'autre part, car pour ces tâches la détection de l'erreur a peu de chance d'arriver lors de relectures de l'énoncé.

Résumons notre pensée en croisant la dimension de conflit cognitif d'une situation ou d'un énoncé avec la dimension sociale de la situation.

Sans aucune communication, sans présence de conflit cognitif, par exemple quand nous avons une intuition très forte sur un sujet, le raisonnement produit des raisons assez faibles et s'en contente. Avec plus de conflits, ou si nous n'avons pas d'intuitions fortes sur un sujet, le raisonnement peut être poussé à évaluer de façon un peu moins laxiste les arguments qu'il produit et à envisager des raisons pour les alternatives. Dans le cas d'étudiants envisageant les raisons pour partir en vacances s'ils n'ont pas leurs examens ou s'ils ont leurs examens. Nous l'avons vu, l'incohérence entre ces raisons pousse les sujets à ne pas choisir. Ou bien dans la condition silence

de notre expérience avec le bonbon et la baguette, les bonnes réponses des voisins peuvent pousser à s'interroger sur les raisons pour cette réponse. Nous avons avancé l'idée que ce conflit pouvait être résolu par certains sujets avec des raisons sur le manque de compétences des autres et ainsi ne pas les faire envisager sérieusement que la réponse pourrait être 0.10€. Dans des cas plus proches de la vie quotidienne, nous pouvons penser que ces cas correspondent, pour les idées politiques par exemple, à des échecs d'attributions de raisons pour des pensées différentes de la nôtre, ou plutôt à l'attribution de raison du type « ils pensent ça parce qu'ils sont égoïstes ou stupides».

Lorsque nous recevons des arguments sans dialogue, les cas sans conflits correspondent à des situations où nous recevons un argument peu convaincant à propos d'un sujet sur lequel nous avons une forte opinion contraire. Lorsque vous recevez un argument pour la réponse « 0.10€ » alors que vous avez la bonne réponse « 0.05€ » par exemple. Les cas avec conflit correspondent eux à la réception d'un argument fort sur un sujet pour lequel nous avons une forte opinion dans le sens inverse. Par exemple, lorsque quelqu'un vous explique pourquoi « oui » est la bonne réponse et non « on ne peut pas savoir » au problème de Linda. Ces cas peuvent aussi correspondre à la réception d'un argument décent pour un sujet sur lequel nous n'avons pas d'opinion forte. Dans ces cas là, nos résultats suggèrent que la qualité des arguments reçus, même sans dialogue, peuvent surmonter une extrême confiance en notre opinion initiale mais également un manque de confiance en la source.

Enfin, dans le cas de dialogue interactif, nos données sur le raisonnement en groupe montrent assez clairement que la présence de conflit au sein du groupe est un facteur très important de ces situations. Si tous les membres du groupe sont déjà d'accord, la discussion a toutes les chances de mener à une polarisation pouvant mener les membres à avoir une grande confiance à partir de très peu d'éléments. C'est évidemment encore pire si la discussion démarre avec des membres déjà tous confiants en la même intuition.

Lorsque différentes opinions sont présentes dans le groupe, en revanche, l'échange d'arguments offre de grands bénéfices permettant aux meilleures raisons d'être conservées. Même dans des cas où il n'y a pas de bonnes réponses claires comme les discussions politiques, certains travaux ont mis en évidence que l'échange d'arguments entre citoyens américains républicains et démocrates par exemple, tend à rendre les gens moins extrêmes dans leurs opinions initiales (Mercier et Landemore 2012).

Nous avons résumé tous les cas possibles entre présence de conflit et situation argumentative dans le tableau suivant :

	PEU OU PAS DE CONFLIT	CONFLIT
 Pas de communication	« On a pas d'information à propos de Linda donc on peut pas savoir ». « Le bonbon fait 10 centimes car avec la baguette ça fait 1€ »	« J'hésite à partir en vacances avant de connaître mes résultats » « …ça paraît quand même bizarre que deux de mes voisins aient répondus 0.05€. ».
 Réception d'arguments sans dialogue	« La réponse n'est pas 0.05€ mais 0.10€ car la baguette vaut 1€ » « Ton équipe va perdre ce soir car cette année ils sont irréguliers »	« Soit Linda est mariée, soit elle ne l'est pas, si elle l'est [...] donc la réponse est oui » Ou, autre cas: « Vu le vent qu'il y a ce soir, il va faire beau demain »
 Dialogue interactif	« J'ai vu sur un site que cet acteur était mort » « Vu son âge ce n'est pas étonnant » « Oui c'est aussi sur un site étranger »	« C'est votre droit d'être fanatique de la peine de mort. Mais c'est notre droit d'hésiter à envoyer un être humain sur la chaise électrique pour un crime qu'il n'a peut-être pas commis. Personne, ici, ne connaît la vérité. Ni vous, ni moi. »

Chapitre 8 – Conclusion

L'objectif de cette thèse est de proposer une comparaison entre d'une part, une vision intellectualiste du raisonnement, considérant le raisonnement comme un mécanisme améliorant la prise de décision individuelle, et d'autre part, une vision interactionniste du raisonnement, faisant l'hypothèse que le raisonnement est avant tout un mécanisme social.

Dans le chapitre 2, nous avons apporté des éléments qui nous permettent de remettre en question l'hypothèse historique associant mécanismes de raisonnement et logique formelle. Nous avons conclu qu'il n'y a pas d'arguments théoriques ou expérimentaux assez forts pour faire l'hypothèse que l'esprit humain implémente des lois logiques, ni même que ces lois représentent une norme pertinente pour la psychologie du raisonnement. Le chapitre 3 a présenté les versions les plus récentes de l'hypothèse intellectualiste, incarnées par les théories à processus duels du raisonnement, aujourd'hui dominantes dans le domaine. En particulier, nous avons discuté des données expérimentales appuyant la théorie commune d'Evans et Stanovich, concluant qu'elles n'étaient pas suffisantes pour conclure à l'existence de deux systèmes cognitifs ni même à deux types de processus. Ce chapitre nous a également permis d'introduire de récents travaux mettant en évidence le fait que les sujets qui donnent la réponse incorrecte intuitive à certaines tâches de raisonnement semblent malgré tout le détecter implicitement.

Dans le chapitre 4, nous avons proposé une alternative aux visions intellectualistes du raisonnement, la théorie argumentative du raisonnement. A partir de considérations évolutionnistes sur les contextes qui auraient pu faire évoluer les mécanismes de raisonnement humain, nous avons présenté une caractérisation du raisonnement remettant en cause non seulement l'idée que la fonction du raisonnement est de corriger nos intuitions, mais également que le raisonnement est défini par opposition aux intuitions. Nous avons, en particulier, introduit la possibilité que le raisonnement soit un mécanisme intuitif spécialisé dans le traitement des raisons dont la fonction principale serait l'argumentation en contexte dialogique.

Nous avons présenté dans le chapitre 5 des données expérimentales mettant en évidence le phénomène de sur-confiance et avons tenté de l'expliquer par la fonction argumentative du raisonnement. Nous avons présenté des données expérimentales montrant l'existence du biais vers son côté, suggérant que plutôt que d'essayer de corriger nos intuitions initiales, le raisonnement chercherait au contraire à les justifier. De plus, nous avons expliqué pourquoi dans l'hypothèse interactionniste, notre raisonnement aurait également tendance en contexte individuel à produire des arguments de faible qualité et à s'en satisfaire. En particulier, nous avons offert une démonstration expérimentale de ce que nous avons appelé la paresse sélective du raisonnement, deuxième trait du raisonnement avec le biais vers son côté permettant d'expliquer le phénomène de sur-confiance.

Enfin dans le chapitre 6, nous avons avancé des résultats montrant la grande supériorité des groupes par rapport aux individus dans la résolution de tâches de raisonnement. Nous avons montré également que même des psychologues spécialistes du raisonnement sous-estiment cette supériorité, puis avons expliqué celle-ci et en avons donné certaines limites.

Pour conclure notre réflexion, résumons les contrastes les plus marqués entre la théorie argumentative et les théories intellectualistes du raisonnement.

D'une part, le contraste avec les théories logicistes est clair. Les théories normatives comme la logique ne sont, de notre point de vue, que peu pertinentes pour comprendre le raisonnement humain. Les sujets n'essayent de s'y conformer qu'au sens où ils veulent apparaître comme des individus cohérents. Lorsque des outils logiques ou mathématiques sont déployés pour argumenter en faveur d'une réponse, les sujets peuvent admettre qu'il serait plus cohérent pour eux d'accepter la bonne réponse, acceptant ainsi la fonction de ces outils comme assurant la cohérence de nos pensées.

D'autre part, le contraste avec les théories à processus duels est également très apparent. L'idée que le raisonnement est un module qui traite intuitivement les liens entre raison et conclusion va clairement à l'encontre de l'idée que le raisonnement repose sur des types de processus différents définis par exemple comme impliquant la mémoire de travail. La théorie argumentative du raisonnement n'offre de liens avec les théories à processus duels que dans le sens où la production et l'évaluation de raisons pourraient être vues comme deux utilisations différentes du même mécanisme. Loin de représenter un système différent des autres, le raisonnement serait un module de plus à notre collection.

Avant de terminer cette thèse en proposant une ouverture sur l'idée d'éducation au raisonnement, reprenons le tableau croisé entre conflit et situation argumentative. A partir des différents cas décrits précédemment, indiquons ceux pour lesquels c'est la production de raisons qui domine (en rouge), et ceux pour lesquels c'est l'évaluation de raisons qui domine (en bleu).

	PEU OU PAS DE CONFLIT	CONFLIT
 Pas de communication	Raisonner seul sur un sujet pour lequel nous avons une opinion/ intuition forte	Raisonner seul sur un sujet sur lequel nous avons des opinions/intuitions conflictuelles
 Réception d'arguments sans dialogue	Recevoir un argument peu concluant sur un sujet pour lequel nous avons une opinion/intuition forte	Recevoir un argument fort sur sur un sujet pour lequel nous avons une opinion/intuition forte Ou Un argument décent sur un sujet pour lequel nous n'avons qu'une faible opinion/intuition
 Dialogue interactif	Echanger des arguments avec d'autres sur un sujet pour lequel tous les interlocuteurs sont d'accord	Echanger des arguments avec d'autres sur un sujet pour lequel les interlocuteurs ne sont pas d'accord

Ce tableau est tiré de Mercier (*in press*) et il permet de résumer l'intégration de la théorie argumentative du raisonnement avec différents domaines de recherche sur le raisonnement. La production domine pour les effets de polarisation et de sur-confiance, en groupe ou seul. Lors de l'échange d'arguments en cas de conflit d'idées ou lors de la réception de très bons arguments contre notre point de vue, l'évaluation domine et permet une forme d'intelligence collective par l'évaluation exigeante des raisons des autres, menant à une sélection des meilleurs arguments. Le cas le plus équilibré entre la production et l'évaluation est celui du raisonnement solitaire avec conflit ou une absence d'intuition forte. Le raisonnement peut nous servir d'outil d'enquête solitaire dans ces cas-là, la production de raisons domine mais une certaine évaluation nous permet d'éviter des actions difficilement justifiables.

Par contraste, lorsque nous avons des intuitions relativement fortes, ou des croyances, des expériences, sur une question, les processus biaisés de production de raisons dominent sur les processus d'évaluation relativement objectifs. Ils mènent dans la majorité des cas sans conflit d'idées, à un phénomène de sur-confiance individuelle et collective.

Ouverture

A partir des différences très nettes entre les théories que nous avons présentées, que ce soit sur la caractérisation qu'elles offrent des mécanismes du raisonnement, la fonction qu'elles lui attribuent ou leur description algorithmique, nous pouvons nous interroger sur les conseils, potentiellement contradictoires, qu'elles peuvent donner à la société. Ouvrons cette thèse avec quelques considérations appliquées, en particulier dans le domaine de l'éducation. Essayons d'imaginer brièvement les propositions de chaque théorie pour entrainer les élèves et étudiants à mieux raisonner.

Dans la vision logiciste, cela consisterait en un entrainement à la pensée abstraite, à la logique formelle et aux mathématiques.

Au vu de l'importance des outils mathématiques dans nos sociétés modernes, ne serait-ce que comme moyen de sélection, nous pouvons concéder qu'un entrainement à la logique pourrait aider à mieux raisonner dans les domaines où ces outils sont prévalents. L'entrainement au maniement de structures abstraites n'est qu'un entrainement à des procédures et des stratégies spécifiques, particulièrement contre-intuitives au départ certes, à cause de l'abstraction des objets sur lesquels elles portent. Espérer plus de bénéfices que la maitrise de ces outils reviendrait à penser que ces structures ont une place spéciale dans notre pensée. On peut au moins en douter.

Dans la vision du raisonnement à processus duels, améliorer son raisonnement pourrait consister à entrainer ses capacités d'inhibition. On peut, par exemple, imaginer qu'entrainer les capacités d'inhibition puisse remédier à des problèmes d'impulsivité. Un

entrainement à l'inhibition a cependant toutes les chances d'être très général et ne pas être spécifique au module du raisonnement dans le sens où nous l'entendons.

Quant à la théorie argumentative du raisonnement, on pourrait penser que la reconnaissance d'arguments fallacieux serait une de ses propositions. Cela serait pourtant une erreur. Des études ont montré qu'il n'y a pas vraiment de forme d'argument fallacieuse en soi. Selon les sujets et les contextes, même des arguments de type ad hominem ou dit « à pente glissante » peuvent être des arguments parfaitement valides (Hahn & Oaksford 2006). Pour ne prendre qu'un exemple, un argument du type « Si cette loi passe, c'est le début de quelque chose de plus dangereux » est un argument classé comme étant d'une forme fallacieuse, mais peut très bien être d'une grande pertinence dans certains contextes. De plus, l'entrainement à la reconnaissance d'arguments fallacieux à toutes les chances de produire les effets inverses que ceux espérés. Les étudiants risquent notamment de repérer ces arguments dans la pensée des autres mais pas forcément dans la leur, menant potentiellement à une polarisation des idées encore plus forte (pour une vision plus complète de ce que pourrait apporter la théorie argumentative dans le domaine de l'éducation, voir Mercier, Boudry, Paglieri et Trouche (in press)).

Si un domaine de recherche a cependant reconnu depuis longtemps les avantages des discussions en groupe c'est bien celui des sciences de l'éducation, développant notamment le domaine de l'apprentissage collaboratif (Voir Slavin 1996 ou Nusbaum 2008 pour des revues). Des études de Kuhn et Crowell (2011) en particulier, suggèrent qu'une pratique régulière de l'argumentation en groupe offre des bénéfices qui vont au-delà des sujets discutés et de l'argumentation elle-même. Les étudiants habitués à l'argumentation tout au long de l'année écrivent, lorsqu'ils raisonnent seuls, des essais plus nuancés, discutant plus souvent de points de vue alternatifs que d'autres étudiants n'ayant pas bénéficié de discussions en groupe.

De plus, nous pouvons penser que dans une société démocratique, les plus grands bénéfices des discussions en groupe dans les milieux éducatifs se trouvent dans la

compréhension des arguments défendant des points de vue opposés à ceux des apprenants.

Au-delà de leur compréhension, c'est surtout l'entraînement à la reconnaissance même de l'existence d'arguments sensés, pour le point de vue des autres, qui pourra créer une culture du « désaccord respectueux » dans les salles de classe et nous faire espérer un monde où les humains raisonneraient mieux.

Bibliographie

Baratgin, J., D. Over, and G. Politzer. “New Psychological Paradigm for Conditionals and General de Finetti Tables.” Mind & Language 29, no. 1 (February 1, 2014): 73–84. doi:10.1111/mila.12042.

Braine, M.D.S. “On the Relation between the Natural Logic of Reasoning and Standart Logic.” Psychological Review 85 (1978): 1–21.

Braine, M.D.S., and D.P. O’Brien. Mental Logic. Mahwah: Lawrence Erlbaum Associates Ltd, 1998.

Brissiaud, Rémi, and Emmanuel Sander. “Arithmetic Word Problem Solving: A Situation Strategy First Framework.” Developmental Science 13, no. 1 (January 1, 2010): 92–107. doi:10.1111/j.1467-7687.2009.00866.x.

Byrne, R. M. “Suppressing Valid Inferences with Conditionals.” Cognition 31, no. 1 (1989): 61–83.

Carruthers, Peter, Stephen Laurence, and Stephen Stich. The Innate Mind: Structure and Contents. Oxford University Press, 2005.

Chaiken, S., and Y. Trope. Dual-Process Theories in Social Psychology. New York: The Guilford Press, 1999.

Corriveau, Kathleen H., and Katelyn E. Kurkul. “‘Why Does Rain Fall?’: Children Prefer to Learn From an Informant Who Uses Noncircular Explanations.” Child Development 85, no. 5 (2014): 1827–35.

Evans, Jonathan St B. T., and Keith E. Stanovich. “Dual-Process Theories of Higher Cognition Advancing the Debate.” Perspectives on Psychological Science 8, no. 3 (May 1, 2013): 223–41. doi:10.1177/1745691612460685.

Evans, J.St.B.T. “Logic and Human Reasoning: An Assessment of the Deduction Paradigm.” Psychological Bulletin 128, no. 6 (2002): 978–96.

Evans, J.St.B.T., J.L. Barston, and P. Pollard. “On the Conflict between Logic and Belief in Syllogistic Reasoning.” Memory and Cognition 11 (1983): 295–306.

Evans, J.St.B.T., and K. Frankish. In Two Minds. Oxford: Oxford University Press, 2009.

Evans, J.St.B.T., S.E. Newstead, and R.M.J. Byrne. Human Reasoning: The Psychology of Deduction. Hove, UK: Lawrence Erlbaum Associates Ltd., 1993.

Evans, J.St.B.T., and D.E. Over. Rationality and Reasoning. Hove: Psychology Press, 1996.

Frederick, S. “Cognitive Reflection and Decision Making.” Journal of Economic Perspectives 19, no. 4 (2005): 25–42.

Gigerenzer, Gerd, and Wolfgang Gaissmaier. “Heuristic Decision Making.” Annual Review of Psychology 62, no. 1 (2011): 451–82. doi:10.1146/annurev-psych-120709-145346.

Girotto, V., M. Kemmelmeier, D. Sperber, and J.-B. Van der Henst. “Inept Reasoners or Pragmatic Virtuosos? Relevance and the Deontic Selection Task.” Cognition 81, no. 2 (2001): 69–76.

Hahn, U., and M. Oaksford. “A Bayesian Approach to Informal Argument Fallacies.” *Synthese* 152, no. 2 (2006): 207–36.

Hall, Lars, Petter Johansson, and Thomas Strandberg. “Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey.” PloS One 7, no. 9 (2012): e45457.

Henle, Mary. “On the Relation between Logic and Thinking.” Psychological Review 69, no. 4 (1962): 366–78. doi:10.1037/h0042043.

Johnson-Laird, P. N. Mental Models. Cambridge, UK: Cambridge University Press, 1983.

Kahneman, D., and S. Frederick. “A Model of Heuristic Judgment.” In The Cambridge Handbook of Thinking and Reasoning, edited by K. Holyoak and R.G. Morrison, 267–94. Cambridge, UK: Cambridge Univiversity Press, 2005.

Keren, Gideon, and Yaacov Schul. “Two Is Not Always Better Than One A Critical Evaluation of Two-System Theories.” Perspectives on Psychological Science 4, no. 6 (November 1, 2009): 533–50. doi:10.1111/j.1745-6924.2009.01164.x.

Khemlani, Sangeet, and P. N. Johnson-Laird. “Theories of the Syllogism: A Meta-Analysis.” Psychological Bulletin 138, no. 3 (May 2012): 427–57. doi:10.1037/a0026841.

Koriat, Asher. “How Do We Know That We Know? The Accessibility Model of the Feeling of Knowing.” Psychological Review 100, no. 4 (1993): 609–39. doi:10.1037/0033-295X.100.4.609.

Kruglanski, Arie W., and Edward Orehek. “Partitioning the Domain of Social Inference: Dual Mode and Systems Models and Their Alternatives.” Annual Review of Psychology 58, no. 1 (2007): 291–316. doi:10.1146/annurev.psych.58.110405.085629.

Kuhn, D. The Skills of Arguments. Cambridge: Cambridge University Press, 1991.

Kuhn, D., and A. Crowell. “Dialogic Argumentation as a Vehicle for Developing Young Adolescents’ Thinking.” Psychological Science 22, no. 4 (2011): 545.

Levesque, H J. “Making Believers out of Computers.” Artif. Intell. 30, no. 1 (October 1986): 81–108. doi:10.1016/0004-3702(86)90068-8.

Lombrozo, Tania. “The Instrumental Value of Explanations.” Philosophy Compass 6, no. 8 (2011): 539–51.

Manktelow, Ken. Reasoning and Thinking. Hove: Psychology Press, 1999.

Mercier, H., and H. Landemore. “Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation.” Political Psychology 33, no. 2 (2012): 243–58.

Mercier, H., and Dan Sperber. “Intuitive and Reflective Inferences.” In In Two Minds, edited by J.St.B.T. Evans and K. Frankish, 149–70. New York: Oxford University Press, 2009.

Mercier, H., and Dan Sperber. “Why Do Humans Reason? Arguments for an Argumentative Theory.” Behavioral and Brain Sciences 34, no. 2 (2011): 57–74.

Mercier, Hugo. “Integrating Reasoning Research with the Argumentative Theory,” In press.

Mercier, Hugo, and Dan Sperber. The Enigma of Reason. In prep, n.d.

Morrison, Robert G. The Oxford Handbook of Thinking and Reasoning. OUP USA, 2012.

Neys, Wim De, Sofie Cromheeke, and Magda Osman. “Biased but in Doubt: Conflict and Decision Confidence.” PLOS ONE 6, no. 1 (January 25, 2011): e15954. doi:10.1371/journal.pone.0015954.

Neys, Wim De, Sandrine Rossi, and Olivier Houdé. “Bats, Balls, and Substitution Sensitivity: Cognitive Misers Are No Happy Fools.” Psychonomic Bulletin & Review 20, no. 2 (February 16, 2013): 269–73. doi:10.3758/s13423-013-0384-5.

Nickerson, R.S. “Confirmation Bias: A Ubiquitous Phenomena in Many Guises.” Review of General Psychology 2 (1998): 175–220.

Nussbaum, E. M. “Collaborative Discourse, Argumentation, and Learning: Preface and Literature Review.” Contemporary Educational Psychology 33, no. 3 (2008): 15.

Oaksford, M., and N. Chater. Bayesian Rationality: The Probabilistic Approach to Human Reasoning. Oxford, UK: Oxford University Press, 2007.

Oaksford, Mike, and Nick Chater. “Theories of Reasoning and the Computational Explanation of Everyday Inference.” Thinking and Reasoning 1, no. 2 (1995): 121–52.

Osman, M. “An Evaluation of Dual-Process Theories of Reasoning.” Psychonomic Bulletin and Review 11, no. 6 (2004): 988–1010.

Perkins, D.N. “Postprimary Education Has Little Impact on Informal Reasoning.” Journal of Educational Psychology 77 (1985): 562–71.

Piaget, J., and B. Inhelder. The Growth of Logical Thinking from Childhood to Adolescence. New York: Basil Books, Inc., 1958.

Posner, M. I., and C. R. R. Snyder. “Attention and Cognitive Control.” In Information Processing and Cognition: The Loyola Symposium, edited by R.L. Solso. Hillsdale, NJ: Erlbaum, 1975.

Reber, A.S. Implicit Learning and Tacit Knowledge. New York: Oxford University Press, 1993.

Rips, L.J. “Cognitive Processes in Propositional Reasoning.” Psychological Review 90, no. 1 (1983): 38–71.

Rips, L. J. The Psychology of Proof: Deductive Reasoning in Human Thinking. Cambridge, MA: MIT Press, 1994.

Shafir, E., and A. Tversky. “Thinking through Uncertainty: Nonconsequential Reasoning and Choice.” Cognitive Psychology 24, no. 4 (1992): 449–74.

Slavin, R.E. “Research on Cooperative Learning and Achievement: What We Know, What We Need to Know.” Contemporary Educational Psychology 21, no. 1 (1996): 43–69.

Sloman, S.A. “The Empirical Case for Two Systems of Reasoning.” Psychological Bulletin 119, no. 1 (1996): 3–22.

Sperber, Dan. “In Defense of Massive Modularity.” In Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler, edited by E. Dupoux, 47–57. Cambridge, Massachusetts: MIT Press, 2001.

Stanovich, K.E. The Robot’s Rebellion. Chicago: Chicago University Press, 2004.

Stanovich, K.E. Who Is Rational?: Studies of Individual Differences in Reasoning. Mahwah, NJ: Lawrence Erlbaum, 1999.

Stanovich, Keith E., and Richard F. West. “Advancing the Rationality Debate.” Behavioral and Brain Sciences 23, no. 05 (October 2000): 701–17. doi:null.

Stanovich, K.E., and R.F. West. “Individual Differences in Reasoning: Implications for the Rationality Debate.” Behavioral and Brain Sciences 23 (2000): 645–726.

Teasdale, John D. “Metacognition, Mindfulness and the Modification of Mood Disorders.” Clinical Psychology & Psychotherapy 6, no. 2 (May 1, 1999): 146–55. doi:10.1002/(SICI)1099-0879(199905)6:2<146::AID-CPP195>3.0.CO;2-E.

Tenenbaum, Joshua B., Thomas L. Griffiths, and Charles Kemp. “Theory-Based Bayesian Models of Inductive Learning and Reasoning.” Trends in Cognitive Sciences 10, no. 7 (2006): 309–18.

Toplak, Maggie E., and Keith E. Stanovich. “The Domain Specificity and Generality of Disjunctive Reasoning: Searching for a Generalizable Critical Thinking Skill.” Journal of Educational Psychology 94, no. 1 (2002): 197–209. doi:10.1037/0022-0663.94.1.197.

Tversky, A., and E. Shafir. “The Disjunction Effect in Choice under Uncertainty.” Psychological Science 3, no. 5 (1992): 305–9.

Vergnaud, Gérard. “A Classification of Cognitive Tasks and Operations of Thought Involved in Addition and Subtraction Problems.” Addition and Subtraction: A Cognitive Perspective, 1982, 39–59.

Wason, P.C. “Reasoning.” In New Horizons in Psychology: I, edited by B.M. Foss, 106–37. Harmandsworth, England: Penguin, 1966.

Wason, P. C. “Reasoning about a Rule.” Quarterly Journal of Experimental Psychology 20, no. 3 (August 1, 1968): 273–81. doi:10.1080/14640746808400161.

Wayne, Carly, Roni Porat, Maya Tamir, and Eran Halperin. “Rationalizing Conflict The Polarizing Role of Accountability in Ideological Decision Making.” Journal of Conflict Resolution, 2015, 0022002714564431.