Reply

Some Clarifications About the Argumentative Theory of Reasoning: A Reply to Santibáñez Yañez (2012).

HUGO MERCIER

Université de Neuchâtel
Centre de sciences cognitives
Espace Louis Agassiz 1 2000 Neuchâtel
Switzerland
hugo.mercier@gmail.com

1. Introduction

In "Mercier and Sperber's Argumentative Theory of Reasoning: From Psychology of Reasoning to Argumentation Studies" (2012) Cristian Santibáñez Yañez offers an interesting take on a new theory of reasoning put forward by Dan Sperber and myself.[1] His comments are especially interesting since they come from the perspective of argumentation studies ("traditionally *dialectics, rhetoric* and *(informal) logic*," 155), a field that Santibáñez Yañez contends has been neglected in this novel theory. After very briefly summarizing the main idea of the argumentative theory of reasoning and clarifying some points for which Santibáñez Yañez may not be offering an entirely accurate representation, the present article will offer a suggestion regarding the potential for mutual enrichment between argumentation studies and the argumentative theory of reasoning.

Several domains—probably most domains—of experimental psychology are dominated by what can be called the *classical view* of reasoning. This perspective posits that the main function of reasoning is to correct misguided intuitions, helping the reasoner reach better beliefs and make better decisions (e.g., Kahneman, 2003; Stanovich, 2004). Theoretical and empirical considerations led Sperber to question the plausibility of this theory (Sperber, 2000, 2001), suggesting instead that the main function of reasoning is to argue: to produce arguments so we can convince others and to evaluate others' arguments so as to

[1] All unattributed quotes are drawn from this article.

© Hugo Mercier. *Informal Logic*, Vol. 32, No. 2 (2012), pp. 259-268.

be convinced only when appropriate. This argumentative theory of reasoning has received empirical support from many domains of psychology (for a brief review, see Mercier, in press-a). Santibáñez Yañez offers a summary of the theory, so this need not be belaboured here. However, two points that may require conceptual clarification in light of Yanez' summary are presently examined.

2. Reasoning and System 2 mechanisms

The argumentative theory of reasoning is partly anchored in *dual process accounts of reasoning* (see Evans, 2008). These accounts generally posit the existence of two types of mental mechanisms. System 1 mechanisms are fast, effortless, unconscious, and prone to systematic biases. System 2 mechanisms are their negative: slow, effortful, conscious and supposedly able to correct System 1's mistakes. Dual process accounts have flourished in different areas of psychology, originally in memory (Schacter, 1987), learning (Berry & Dienes, 1993; Reber, 1993) and attention (Posner & Snyder, 1975), more recently in social psychology (Chaiken & Trope, 1999), reasoning (Evans & Frankish, 2009) and judgment and decision making (Kahneman, 2003). Given the many different perspectives converging on dual process accounts, it may not be entirely surprising, as Santibáñez Yañez notes, that "there is no clear agreement on what characterizes each system, how they are related, and which processes and functions are inherent to each" (p. 136).

While this assessment is hardly disputable, it is unclear whether that constitutes an indictment of the argumentative theory of reasoning, since one of its potential strengths is precisely to propose a more precise definition of reasoning—into which Santibáñez Yañez delves in some detail (p. 136ff). Indeed, as they now stand, it is dubious whether the argumentative theory of reasoning and typical dual process accounts have much in common beyond the initial—but crucial—insight that a few mental mechanisms exhibit traits that are substantially different from the rest of our cognitive apparatus. To put it as succinctly as possible, the argumentative theory of reasoning sees reasoning as a mechanism that finds and evaluates reasons. As such, it is part of the family of metarepresentational mechanisms: mechanisms that deal with representations of representations. In the case of reasoning, the representations represented are premises and conclusions and what is represented is whether a given premise is a good reason to accept a given conclusion.

There are at least two problems with the way reasoning and System 2 tend to be associated, explaining why the argu-

mentative theory differs considerably from most dual process accounts. (1) System 2 is far from being restricted to reasoning and (2) reasoning has many traits generally attributed to System 1 processes. On point (1), not only do several mechanisms beyond reasoning belong to what is typically referred to as System 2, such as thinking or planning (see Mercier & Sperber, 2011a), but many other cognitive mechanisms can be recruited to function in a "System 2 way." For instance, when you're looking for someone in a crowd, you recruit a typical System 1 mechanism (face recognition) in a slow and effortful manner more characteristic of System 2 processes.

Regarding point (2), in the right contexts reasoning shares many traits typically associated with System 1 processes. When people argue, they generally find and evaluate arguments quickly and effortlessly (Mercier & Sperber, 2011b). Moreover, reasoning shares an even more important trait with System 1 processes: it has an important central unconscious dimension, which tends to be neglected. Reasoning relies on intuitions about reasons: whether a given premise is a good reason to accept a given conclusion. People usually have little conscious access into why they have these intuitions (e.g., why most people think that arguments of the form "If *p* then *q*; *p*; therefore *q*" or that "I think therefore I am" are good arguments).

While the argumentative theory of reasoning substantially differs from most dual process accounts, it is somewhat closer to Stanovich's account. Because it defines reasoning as a metarepresentational mechanism, the argumentative theory of reasoning shares with Stanovich's theory the idea that "decoupling" (essentially a synonym of metarepresentating here) is an important trait of System 2 mechanisms—a point on which Santibáñez Yañez opines (p. 147). It should be stressed, however, that several other cognitive mechanisms rely on "decoupling," most notably, Theory of Mind (the attribution of mental states to others) and pragmatics (the attribution of speaker's meaning). Not only do these other mechanisms typically function quickly and effortlessly, they are also often blamed for reasoning's failures (including in Stanovich's own account, 2004; see also Levinson, 1995). As a result, it is important to distinguish reasoning *per se* from other metarepresentational mechanisms, a distinction that seems better marked in the argumentative theory of reasoning than in Stanovich's account.

3. The benefits of reasoning and argumentation

Santibáñez Yañez presents a rather bleak picture of what the benefits of reasoning and argumentation are supposed to be in the argumentative theory of reasoning: "argumentation is repeatedly presented as a dimension that does not improve cognitive skills and only as a side-effect provides some gains for individuals" (p. 135); "what this theory challenges is that humans make good decisions, maintaining that we prefer to make decisions we can justify more easily in front of others" (p. 139). It is true that articles reviewing the empirical support for the argumentative theory of reasoning have tended to dwell on reasoning's supposed failures. The main reason for this, however, is that it is this type of evidence that provides the strongest support for the argumentative theory of reasoning against the classical view. For instance, a general observation is that people perform poorly in reasoning tasks individually but improve when reasoning in groups. On its own, the good performance of reasoning in groups is not a very strong argument in favor of the argumentative theory of reasoning and against the classical view—the classical view does not predict that reasoning should work poorly in group settings. By contrast, the fact that people perform poorly on reasoning tasks individually is predicted by the argumentative theory of reasoning while going against the prediction of the classical view. It makes sense, therefore, that the focus should have been on the poor performance of the lone reasoner or, rather, on the contrast between these poor performance and the good performance in group settings.

The general picture of reasoning painted by the argumentative theory of reasoning should most emphatically *not* be a bleak one. The theory turns a deeply flawed individual mechanism into a wonderfully designed argumentative device. It predicts that when reasoning is used in the proper circumstances—among people who disagree but are ready to change their mind when confronted with good arguments—it can produce considerable epistemic benefits. More specifically, what makes group discussion a propitious context for reasoning to yield epistemic improvements is the back and forth between the positions of producer and evaluator of argument.

Santibáñez Yañez rightfully stresses this back and forth, but he uses it as an argument against the idea that people use mechanisms of epistemic vigilance to evaluate communicated information: "In other words: in argumentative scenarios, to ask for clarification, to counter-argue, or to put forward doubts are more than passive mechanisms, which the simple idea of vigilance seems to convey" (pp. 142-3). Such episodes, however, do not argue against the idea of epistemic vigilance. In the dynamic of a conversation, it is often the case that when

someone rejects a statement, she will be expected either to justify her rejection or to convince the interlocutor that he is the one who should change his mind. But people still need a mechanism of epistemic vigilance to know what statements they should reject, or at least maintain a temporary stance of doubt towards, which is a necessary precondition if they are to ask for clarifications or to counter-argue.

Why is the evaluation of arguments so critical for reasoning to yield epistemic improvements? When people *produce* arguments, reasoning exhibits a strong confirmation bias (Mercier, in prep; Nickerson, 1998). This is only to be expected if the goal of reasoning is to convince an audience. Since these arguments are biased, on their own they are apt to lead to epistemic distortions such as belief polarization (Tesser & Conlee, 1975) or overconfidence (Koriat, Lichtenstein, & Fischhoff, 1980). By contrast, when people *evaluate* arguments, their primary goal should be to decide if the argument is good enough to warrant changing their mind about the conclusion. Reasoning itself should try to approximate as best as possible a fair assessment of the argument: after all, failing to accept valuable information is costly. However, other mechanisms may be more reluctant than reasoning to accept that one should change one's mind. Santibáñez Yañez mentions the possibility that people can fail to change their mind when they should, but attributes it—mistakenly, I surmise—to the confirmation bias: "If a speaker, as a natural tendency, and even as a first reaction after getting the answer of the audience, persists in its confirmation bias error, then how is her detection system working?" (p. 143). If someone has a tendency to reject valuable information that clashes with her beliefs, this does not mean that reasoning is at fault. When people have a strong commitment to a point of view, they are likely to reject information that contradicts this point of view *without reasoning*. In such situations, the confirmation bias is not to blame, but instead mechanisms designed to maintain consistency so as not to be thought of as someone who tends to be wrong or a flip-flopper (see Mercier, in press-b).

The distinction between the goals of reasoning when it produces and evaluates arguments may also provide the answer to another challenge set up by Santibáñez Yañez. Santibáñez Yañez refers to the theoretical and empirical work that has applied the concept of Bayesian rationality to argumentation (for review, see Hahn & Oaksford, 2007). This research has shown that people can evaluate arguments in a way that follows the principles of Bayesian rationality. For Yanez, "[t]his idea clearly is contrary to M&S's message that maintains that individual performances are tied to poor outcomes. M&S also claim,

contrary to what Bayesian rationality seems to indicate, that people exhibit confirmation bias all the time without any possibility of choosing alternatives arguments. It is quite surprising that while using more or less the same data the two approaches reach opposite conclusions" (p. 147). As far as I know, however, the Bayesian approach to argumentation has focused on argument evaluation. The argumentative theory of reasoning predicts that argument evaluation should be as unbiased as possible, thereby making essentially the same predictions as the Bayesian approach in this case. By contrast, the presence of a confirmation bias in argument production needs not go against the principles of Bayesian rationality. If the goal of argument production is not epistemic improvement but conviction of an audience, then a rational analysis—Bayesian or of any other type—of this mechanism would also, presumably, predict the existence of a confirmation bias.

As mentioned above, a substantial amount of evidence demonstrating the potentially dire consequences of individual reasoning has been reviewed to support the argumentative theory of reasoning. This does not mean, however, that individual reasoning always leads to poor outcomes. Again, the balance in the exposition is tilted in this direction for argumentative reasons, as I presently explain using the example of reason-based choice.

Psychologists studying judgment and decision making have observed that people sometimes make decisions because reasons supporting these decisions are more accessible than reasons supporting other options (Shafir, Simonson, & Tversky, 1993). Many studies (reviewed in Mercier & Sperber, 2011b) show that this process regularly leads to decisions that are easier to justify, but inferior to some alternatives. But this process will also lead, in many cases, to decisions that are easy to justify *and* good. After all, a decision that can be justified is one that is likely to be deemed good by other people. And people are more often right than wrong. As a result, being careful to make justifiable decisions should often bring about positive outcomes.

Why stress the poor outcomes, then, at the risk of presenting an unduly bleak picture of the usefulness of reasoning? First, poor outcomes are overrepresented in the literature, partly for sociological reasons that render them easier to publish (see Kruger & Savitsky, 2004). Second, these outcomes are the only ones able to test the opposing predictions of the argumentative theory of reasoning and the classical view. According to the argumentative theory, when people reason about a decision in the absence of strong intuitions favoring an option, reasoning will drive them towards the option they can most easily justify, whether it is the best or not. The classical

view, by contrast, should predict that reasoning drives people toward better decisions, period. To test these predictions, one cannot examine cases in which the good decision and the justifiable decision are the same. Hence the stress on decisions that are easy to justify but poor.

Overall, the message of the argumentative theory should be mostly positive. The known flaws of reasoning are reinterpreted as sound design features. Instead of despairing over reasoning's supposed limitations, we should rejoice in the ease with which arguing can turn them into strengths, making good reasoners of us all.

4. The argumentative theory of reasoning and argumentation studies

As Santibáñez Yañez notes, several questions playing an important role in argumentation studies have not been broached within the framework of the argumentative theory of reasoning. Hopefully, the future will bring work filling this gap and, more generally, work drawing both from argumentation studies and cognitive psychology (a forthcoming issue of *Thinking and Reasoning* on the topic of Argumentation, edited Ulrike Hahn and Jos Hornikx is a good step in this direction). Certainly, argumentation studies have much to bring to the psychology of reasoning. By studying debates and discussions in ecological settings, argumentation scholars have become interested in a wide range of arguments. These arguments extend well beyond formal logic, which has been the main focus of the psychology of reasoning. Argumentation studies have also stressed the interplay of different means of persuasion—classically logos, ethos and pathos—something that has not been sufficiently studied in the psychology of reasoning. The argumentative theory of reasoning has drawn heavily from the psychology of reasoning, and it is a cognitive, naturalistic theory. However, its aim is not to deny the complexity of argumentation but rather to provide new tools to better understand it.

One area in which the interplay of argumentation studies and the argumentative theory of reasoning may be especially pregnant with possibilities is that of the categorization of arguments. The categorization of arguments in various arguments schemes is an important topic for argumentation studies (e.g. Walton, Reed, & Macagno, 2008). While the goal of this categorization has been in large part normative—deciding, for instance, what critical questions should be asked to evaluate an argument—one can also wonder about whether laypeople rely on such schemes, however implicitly, when they evaluate argu-

ments. For instance, what are the psychological processes at play when people are confronted with an argument from expertise? Do they have a specific mechanism for this type of argument as opposed to, say, *ad hominem* arguments? According to the argumentative theory of reasoning, people recruit intuitions when they have to evaluate arguments. This can most easily be seen when there is a rather direct relation between an intuition and the argument that recruits it. For instance, if the milk in your bottle smells bad, you'll intuitively know not to drink it. The same intuition is recruited when you evaluate the following argument: "You shouldn't drink this milk because it smells bad." The same process can be applied to arguments from expertise. We are endowed with specialized mechanisms that calibrate our trust in different people as a function of their competence and benevolence, and these judgments intuitively affect the way we evaluate communicated information (Sperber et al., 2010). Arguments from expertise recruit these intuitive processes. It is often easy to see what intuitive process is recruited in a given argument scheme. We intuitively dislike people who do not act in a way that is coherent with their stated beliefs, an intuition recruited in the circumstantial *ad hominem*. We intuitively take into account the number of people who hold an opinion when we evaluate it, an intuition recruited in the *ad populum*. While the link between various intuitive cognitive mechanisms and argument schemes is speculative at this point, it provides ground for both theoretical and empirical work at the junction of argumentation studies and cognitive psychology.

5. Conclusion

Undoubtedly, much work remains to be done in order to bridge argumentation studies and cognitive psychology. By trying to alert cognitive psychologists to the importance of argumentation, the argumentative theory of reasoning will hopefully be seen as a step in this direction.

References

Berry, D.C. & Dienes, Z. (1993). *Implicit learning*. Hove: Erlbaum.

Chaiken, S. & Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. New York: The Guilford Press.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby

(Eds.), *The Adapted Mind* pp. 19–136. Oxford: Oxford University Press.

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. & Frankish, K. (2009). *In Two Minds*. Oxford: Oxford University 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.

Koriat, A., Lichtenstein, S. & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory and Cognition*, *6*: 107–118.

Kruger, J. & Savitsky, K. (2004). The "reign of error" in social psychology: On the real versus imagined consequences of problem-focused research. *Behavioral and Brain Sciences*, *27*(03): 349–350.

Levinson, S.C. (1995). Interactional biases in human thinking. *Social intelligence and interaction*, 221–260.

Mercier, H. (in press-a). Using evolutionary thinking to cut across disciplines: The example of the argumentative theory of reasoning. In T. Zentall & P. Crowley (Eds.), *Comparative Decision Making*. New York: Oxford University Press.

Mercier, H. (in prep). Explaining the confirmation bias.

Mercier, H. (in press-b). The social functions of explicit coherence evaluation. *Mind & Society*.

Mercier, H., & Sperber, D. (2011a). Argumentation: its adaptiveness and efficacy. *Behavioral and Brain Sciences*, *34*(2): 94–111.

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

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

Posner, M.I. & Snyder, C.R.R. (1975). Attention and cognitive control. In R.L. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium*. Hillsdale, NJ: Erlbaum.

Reber, A.S. (1993). *Implicit Learning and Tacit Knowledge*. New York: Oxford University Press.

Santibáñez Yáñez, C. (2012). Mercier and Sperber's Argumentative Theory of Reasoning: From Psychology of Reasoning to Argumentation Studies. *Informal Logic*, *32*(1), 132–159.

Schacter, D.L. (1987). Implicit Memory: History and Current Status. *Journal of experimental psychology. Learning, memory, and cognition*, *13*(3): 501–518.

Shafir, E., Simonson, I. & Tversky, A. (1993). Reason-based choice. *Cognition*, *49*(1-2): 11–36.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.), *Metarepresentations: A Multidisciplinary Perspective* (pp. 117–137). Oxford: Oxford University Press.

Sperber, D. (2001). An evolutionary perspective on testimony and argumentation. *Philosophical Topics*, *29*: 401–413.

Sperber, D., 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.

Tesser, A., & Conlee, M. C. (1975). Some effects of time and thought on attitude polarization. *Journal of Personality and Social Psychology*, *31*(2): 262–270.

Walton, D.N., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge: Cambridge University Press.