Review of Mercier and Sperber's *The Enigma of Reason* (preprint of article to appear in *Teorema*)

Salvador Mascarenhas*
Ecole Normale Supérieure
Département d'Etudes Cognitives
Institut Jean-Nicod
ENS, EHESS, PSL University, CNRS

October 2018

Abstract

Mercier and Sperber argue very convincingly that the traditional intellectualist approach is altogether inadequate to explain the workings of deliberate reasoning about inferences. However, they indict an entire field of inquiry and whole classes of theories of human inferential behavior on this charge, when in fact many of these theories were primarily designed to account for intuitive, non-deliberative, first-order inference making. This tension can be resolved in a constructive way that can propel the field forward into an appropriately ambitious psychology of reasoning: non-interactionist theories of reasoning need to scour their extant empirical coverage for aspects of phenomena that might best be seen as instances of reasoning about reasons, and address those within frameworks that provide satisfactory answers to the challenges raised by Mercier and Sperber. In the process, one's non-interactionist theory of reasoning itself can only become clearer, and will possibly acquire greater explanatory adequacy.

The enigma of reasons explained

In their (2017) book "The Enigma of Reason," Mercier and Sperber (M&S) present a view of reason and a program for its study that addresses a number of old and new challenges. Center stage is a careful discussion of the failures and failings of the traditional approach to reasoning, dubbed the intellectualist view by the authors, which holds that the purpose of the human capacity for reasoning is to track truth and to allow for good decision making. Mercier and Sperber show that this view is ill equipped to provide an explanatory account of human reason given what we know about the failures and successes in how humans deploy and evaluate reasons. If reason aims at truth, why

*The author's work is funded by Agence Nationale de Recherche's grants ANR-17-EURE-0017 and ANR-18-CE28-0008.

are we prey to so many biases, why do we confabulate reasons so promptly, why do we make so many mistakes when reasoning by ourselves? And what of the successes of reason, how does the goal of truth tracking explain that, in social contexts, our capacity for using and evaluating reasons in argumentation so often leads us to good or good-enough decisions, or allows for coordinated action on subjects of great complexity? In response to these challenges, and a host of others, Mercier and Sperber propose an interactionist view of reason: the purpose of this human faculty is to produce reasons to justify oneself, and to produce arguments to convince others.

In what follows, I argue that Mercier and Sperber's work presents a compelling view of a particular phenomenon that can be called "reason" or "reasoning," namely the production and evaluation of reasons. But the psychology of reasoning has at least for the past several decades been about far more than *reasons*, and M&S's sweeping criticism of this work is not entirely justified. In fact, the traditional approach is a valid way to investigate another human phenomenon that can intelligibly be called "reason" and "reasoning." And even if it is not, the interactionist approach fares no better as a foundation for the study of this other kind of reasoning. I conclude that, once we evaluate the theories in the field under this light, M&S's approach is an important advance toward a full understanding of human reasoning, it carves out a chunk of the phenomenon and shows that it is significantly different from the rest, and it lays the foundations for how to understand it. But this work does not offer a broad indictment of the research on reasoning that precedes it.

Inferences, reasons, and inferences about reasons

In my view, Mercier and Sperber's critical outlook on extant theories of reasoning is more negative than what is warranted. Mercier and Sperber are happy to grant that what they call the intellectualist approach has produced a wealth of empirical results on failures and successes of reasoning. After all, a carefully mapped empirical landscape can be of great use to science, even if its cartographers were theoretically misguided.

More importantly though, extant approaches have produced a wealth of *models* of those successes and failures of reasoning. The empirical scope, theoretic insight, and formal rigor of those models vary greatly, from spectacularly sophisticated to modest on all three fronts. But before we scratch the past fifty years of modeling proposals about human reasoning, it is worth checking that all we're left with after salvaging the empirical discoveries is indeed bathwater.

First and higher order inferences

Mercier and Sperber point out that, while *reason* might well be an exclusively human faculty, its super-category *inference making* certainly is not. Indeed, humans and other animals make inferences all the time, say about what to eat, whom to mate with, when to attack, when to flee. These inferences seem to be the product of specialized modules

rather than of some general-purpose faculty.[1] I will call these kinds of inferences *first-order inferences*. They contrast with *reason* as M&S propose we see it: a domain-general faculty that operates on representations of inferences to produce reasons connected to those inferences. Reasons in turn are themselves representational entities that *explain* and *justify* inferences (and presumably all sorts of other things, like desires and actions).

In this discussion, it will be useful to call the product of reason in M&S's sense *higher-order inferences*. One does this not without some abuse of terminology, for they are inferences about *representations* of inferences. But there is a gain in perspicuity: this terminology fulfills M&S's recommendation not to see the reason module as something qualitatively distinct from the lower inferential modules, while pinpointing the difference in levels between the two kinds of inferences. Both first-order inferences and higher-order inferences are the result of modules that deliver intuitive inferences, leading us to expect that they should share some psychological signatures. But we also expect them to differ in interesting ways due to the fact that the principal objects they operate on are of different kinds: precepts and knowledge of the world vs. representations of inferences.

Mercier and Sperber's central criticism of the state of the art in the study of reasoning can be articulated as follows. Most if not all extant theories of reasoning hold, be it overtly or tacitly, that the functional aim of the higher-order inference module is to track truth closely and to achieve better decisions. Yet this idea is at odds with over fifty years of research showing the fallibility of this higher-order inference module at guaranteeing truth tracking and optimal decision making. When humans introspect about reasons, they often end up with arguments that violate fundamental laws of logic. Moreover, humans confabulate reasons that often have little or nothing to do with the actual causes of their lower-level inferential behavior. This is baffling from a view that holds that the capacity to draw higher-order inferences is about truth tracking and good decision making, but it becomes natural and clear once one moves to the idea of social interaction as the functional aim of higher-order inferences.

Stated in this way, I am persuaded by the arguments put forth in the book in favor of this thesis.

The object of study of reasoning

The issue is that it is by no means obvious just how many (and which ones) of the extant theories of reasoning *are about* higher-order inferences. In fact, it is clear that many of them, including the most well-known ones, have instead focused on particular subclasses of first-order inferences. Consider the famous bat-and-the-ball problem, part of the cognitive-reflection test (Frederick 2005).

> A bat and a ball cost $1.10 together. The bat costs $1.00 more than the ball.
> **How much does the ball cost?**

Many people answer that the ball costs $0.10, when in fact $0.05 is the only solution to this simple system of equations. This is a striking result, but what exactly makes this

[1] Like M&S, I use the term "module" to mean roughly "cognitive function."

problem and others like it informative and insightful from a theoretical standpoint?

Sure enough, interesting things happen when we tell participants in this experiment that in fact ten cents is not the right answer. Subjects might engage in what looks like mathematical reasoning: well the bat is one-ten and it's one dollar more than the ball, one-ten minus one equals ten cents. Then, with a bit of prodding, and in particular when confronted with the fact that $1.1 + 0.1 = 1.2$, subjects will often see the light and correct their answer. There are good questions about higher-order inferences to be asked here, and M&S's approach to reason offers a promising framework in which to pursue those questions. But there is also something else going on.

You may not have uttered the incorrect response the first time you were presented with this problem. But almost certainly you had the experience of hearing an insistent voice in your head whispering "stop thinking, this is very easy, the answer is ten cents." As far as you can tell, this voice addressed you prior to any reasoning about the reasons why the answer provided by the voice was right.

This *first-order inference* is also a worthy object of study. Presumably it comes from one of those intuitive modules that deliver low-level inferences, and perhaps M&S are right that to call that "reasoning" is not the most vernacular or conceptually tidy use of the English word. What is clear is that a good chunk of the field has been focusing on precisely these inferences, has been happily calling them "reasoning," and has been developing complex models to describe their workings, the best of which have strong predictive power or theoretic insight.

Two examples: the conjunction fallacy and the Wason selection task

The conjunction fallacy (Tversky and Kahneman 1983) offers an even more instructive example, because we have a large array of competing theories of it.[2] At around 85% rate of conjunctive responses in the original studies, the conjunction fallacy looks a lot like the bat-and-the-ball problem, in that participants feel a strong attraction toward the purportedly fallacious answer.[3] As far as I know, there is good reason to think that the intuitive pull of the conjunctive option precedes any inferences experimental subjects may make about the reasons they have for picking that option. Accordingly, Tversky and Kahneman's account of the phenomenon in terms of representativeness is an account about an entirely unconscious and automatic process that does not require deliberation. Specifically, subjects substitute a question about representativeness (typicality in this case) for a question about probabilities. Notice that nothing in this account is in any way

[2] I agree with M&S that it is suspicious that after more than 50 years of research on reasoning we've achieved so little in way of consensus, a good example of which is the conjunction fallacy. But I believe that the right diagnosis involves much more than just a confusion about first order and higher-order inferences. In my view, the lack of systematic and appropriately sophisticated research on the interplay between interpretation and reasoning has introduced a host of confounds into the field, which we are only now addressing with the appropriate linguistic tools (Mascarenhas 2014).

[3] I say "purportedly" as a nod to the tradition in the psychology of reasoning, behavioral economics, and more recently linguistics of seeking absolving interpretations of the conjunction fallacy that dispel it as an experimental artifact. See for example the work of Hertwig and Gigerenzer (1999) or Dulany and Hilton (1991).

about attaching reasons to representations of inferences. Tversky and Kahneman's story is about the under-the-hood process that delivers the intuitive and attractive first-order response, irrespective of what experimental subjects might want to tell us about the reasons they think they had for picking it.

Now, M&S can of course say that Tversky and Kahneman were looking at something far less interesting than what M&S call reason. But it makes little sense to discard their theory simply on the grounds that if it is not about higher-order inferences then a theory cannot use the English word "reason" and its morphological cousins when it tells us what it is meant to be about.

Consider now the Wason selection task (Wason 1968) as an example where first-order inferences play a more dubious role. Unlike the conjunction fallacy, typical answers to the Wason selection task in its original formulation cover an appreciably wider spectrum, with the most popular answer (turn around antecedent-verifying and consequent-verifying cards, modulo negations) corresponding to about half of responses. Using an old fashioned but often useful tool, introspection suggests to me that I have no immediate pull toward any answers in the Wason selection task. In fact, I feel puzzlement and frustration at the question, and then start engaging in overt deliberation. The fallacious response, it seems to me, comes *alongside* the higher-order reasoning I engage in.

The usefulness of the old ways

Mercier and Sperber argue convincingly that the functional aim of *what they call* "reason" and "reasoning" is not to arrive at good decisions given a particular notion of what is ecologically relevant, but rather to serve as a tool for social interaction.[4] But the old-fashioned dogma is a perfectly plausible hypothesis for the functional aim of the low-level intuitive inferential modules whose existence M&S happily grant.

Indeed, once we look at the arguments against the intellectualist approach through this lens, it becomes clear that they are nowhere near as strong when applied exclusively to first-order, intuitive inferences. Interactionist considerations that are compelling explanations of logically puzzling behavior in argumentation are of little use in the realm of low-level inferences. How is, say, the availability heuristic of Tversky and Kahneman (1974) explained in terms of its usefulness in social interaction?[5] The traditional view that these kinds of processes aim at truth tracking and good decision making certainly still has important challenges to answer, but it is clear that those challenges are not dispelled by an interactionist approach.[6]

[4]There is to my mind an unresolved tension between this position and the authors' observations about epistemic vigilance. Reasoning exists *for* social interaction, yet this is less transparent in the case of hearers than in the case of speakers. Addressees of argumentative discourse have a strong interest in exercising epistemic vigilance, but it is unclear whether and how M&S expect this to fall from the general view of reasoning as aiming at social interactions.

[5]To use the availability heuristic is to answer questions about the frequency of an event in terms of how easily past instances of that event present themselves to one's mind.

[6]Moreover, understanding the truth-tracking failures of a system that aims at truth tracking has always been a crucial goal of the traditional perspective. The heuristics and biases program tells us that we evolved imperfect but approximative strategies meant to give us responses within actionable time frames. Oaksford

Other elements of M&S's arguments in favor of interactionism are less easy to apply to first-order inferences. Higher-order reasoning improves in dialogical contexts, and better decisions can be achieved in such contexts, under certain conditions. But have we reason to think that first-order intuitions themselves improve in dialogical situations? If you solve the bat-and-the-ball problem in conversation with your friends you are more likely to find the solution, but does something deep happen to the little voice in your head that whispers "ten cents"? Is it silenced in a way that is interestingly predicated on the dialogical activity? These questions deserve study of course, but it is as of now entirely plausible to think that they are answered in the negative.

Synthesis

Mercier and Sperber's bleak outlook on state-of-the-art research on reasoning is justified only insofar as the state of the art is meant to be *about* what M&S call "reason" and "reasoning." For example, work from the heuristics and biases paradigm on the conjunction fallacy is very plausibly not about M&S's "reason" and "reasoning" (higher-order inferences), but about intuitive (first order) inferences. As far as I can tell, M&S do not provide arguments impugning research on intuitive inferences from a non-interactionist perspective, and consequently a sizable portion of the existing work on reasoning, conceived in this strict way, is in principle perfectly compatible with M&S's framework. It could well be that the first-order inferential module responsible for the attractiveness of the conjunctive option in the conjunction fallacy is indeed aiming at a judgment of typicality, as Tversky and Kahneman argue. And this substitution of questions is a plausible adaptation under the functional goal of delivering good-enough decisions within actionable time frames.

I propose we do not think of M&S's work as an *alternative* to the past 50 years of theoretical research on reasoning. Instead, let us see it as an important and overdue piece of scholarship that zooms out of low-level accounts of particular classes of fallacies to offer a framework in which every extant theory of reasoning should now situate itself. In particular, the lack of clarity in the field so far between first and higher-order inferences is real and has certainly been pernicious. Thanks to M&S's work, we now see that, despite all the properties the two kinds of inferences share, a unified account in terms of their functional aims is not forthcoming: higher-order inferences most plausibly exist for social interaction, first-order inferences for decision making.

It is interesting in this connection to consider early work on dual-process approaches to reasoning. In particular, Wason and Evans (1975) outline a framework for the study of reasoning that recognizes two kinds of processes: (1) "the processes underlying the

and Chater (2007) argue that in fact our intuitive responses are far more rational than they seem. Philipp Koralus and I toss another brand of fuel into the flame in our work on the erotetics of reasoning (Koralus and Mascarenhas 2013): the contents of humans' propositional thoughts are structured in singular or multiple *alternatives*, a way of recruiting attention in reasoning; but representing multiple alternatives is costly, so we sometimes hastily discard alternatives that evidence suggests are less relevant, producing illusory inferences.

reasoning performance," and (2) "introspective accounts of [reasoning] performance."[7] These two levels correspond quite neatly to the distinction M&S and I make between first order and higher-order inferences. Moreover, there is good reason to think that Wason and Evans consider both processes to be legitimate objects of study for a psychology of reasoning. After all, the authors argue there and elsewhere for the existence of a *matching bias* in the Wason selection task, whereby reasoners solve the task partly by looking for cards that *match* the cards mentioned in the conditional sentence. Since they give precisely this matching bias as an example of a type 1 process, we can only conclude that their account of the Wason selection task involves an account in terms of first-order inferences. Thus, the view that I am suggesting here, where both processes are worthy of study and need to be incorporated into a larger view of reasoning, is by no means novel.

Importantly, this perspective has concrete consequences for how the field should proceed. For one thing, we need to find ways of operationalizing the question of whether a particular class of inferential behavior falls under first-order or higher-order reasoning. If one's theory makes crucial appeal say to rational analysis in purely non-interactionist settings, one must either focus on first-order reasoning or respond satisfactorily to the challenges raised in M&S's work. These questions will be extremely complex in many cases. In the Wason selection task for example, it may well be that we find a conspiracy of low-level processes such as matching bias and high-level reasoning about reasons, which will lead us to error in non-dialogical situations.

We also need to investigate the properties of these low-level inferential modules. Are there several such modules? Do they ever pull in different directions and if so how are disagreements resolved? Is a unified account of their properties at all possible? Since they do not depend on human-specific meta-representational abilities, we expect to find some of these low-level inferential modules in our closest relatives in the animal kingdom. Can those investigations lead to a better evolutionary account of this component of human reasoning?

In conclusion, M&S's thought-provoking book can be seen as offering a broader, conceptually tidier, and more ambitious framework for thinking about human reasoning than the field has seen so far. But there is no good argument in it to reject non-interactionist approaches as a matter of principle, they are in fact very plausibly the right way to approach first-order inferences. A true synthesis of the two lines of research will require work from interactionists and intellectualists alike, but the potential for advance in the field is impressive. Who knows just to what extent otherwise excellent theories of first-order inferences have been led astray by trying to account for instances of higher-order reasoning in the same fashion?

[7] Mercier and Sperber discuss Wason and Evans's work in chapter 2, but they do so from their perspective that only the introspective accounts of performance are "reasoning proper." This is precisely the unwarranted move I criticize in this review.

References

Dulany, D. E., & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, *9*(1), 85–110.

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

Hertwig, R., & Gigerenzer, G. (1999). The conjunction fallacy revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–305.

Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: Bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*, *27*, 312–365.

Mascarenhas, S. (2014). *Formal semantics and the psychology of reasoning: Building new bridges and investigating interactions* (PhD thesis). New York University.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press Cambridge, MA.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Wason, P. C., & Evans, J. S. B. T. (1975). Dual processes in reasoning? *Cognition*, *3*(2), 141–154.