Why Reason? Hugo Mercier’s and Dan Sperber’s The Enigma of Reason: A New Theory of Human Understanding

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Mind &amp; Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>Draft</td>
</tr>
<tr>
<td>Wiley - Manuscript type:</td>
<td>Solicited Review</td>
</tr>
<tr>
<td>Keywords:</td>
<td>metarepresentation, modular theory of mind, evolution of reasoning, evolution of metarepresentation, Dan Sperber, Hugo Mercier</td>
</tr>
</tbody>
</table>

Abstract: The standard view of the function of reason is that it emerged to enable individuals to make better judgements and choices. Once individuals could think better, and once we had suitable communicative tools, individual reasoning acquired a public face; we reasoned together as well as privately, in our own mind. Hugo Mercier and Dan Sperber argue that this gets the story the wrong way around: reasoning evolved for public purposes: to persuade, negotiate, assess. Once it was established publically, perhaps it acquired a private function too. With the exception of a few minor complaints, this evolutionary case is well made. However, Mercier and Sperber embed their evolutionary case within a modular view of the mind, and suggest a modular view of public reasoning itself. So while I find the evolutionary case persuasive, and am sceptical of the cognitive science framework.

Kim Sterelny
Philosophy, ANU

1. Inside-Out, or Outside-In?

One very routine feature of human social life is that we try to persuade one another of our preferred views and courses of action, and that we discuss practical and intellectual problems, trying to solve them collectively. Moreover, when others try to persuade us, we assess what they say, taking into account both our overall view of their reliability and the specifics of the particular proposal they want us to accept. In the public space of interpersonal communication, we give reasons, search for reasons in joint problem solving, and assess the supposed reasons of others. As Mercier and Sperber see it, on the standard picture, one that they shall oppose, this public use of reason is a by-product of private reasoning: the capacity of individuals to work through problems — what to believe, what to do — in their head. Conscious thought plus language enables us to make our conscious thinking public, perhaps improving our decision making by taking advantage of others’ insights and knowledge. But the core capacity — conscious directed thought, directed on the questions of what to believe and do — came first, making interpersonal reasoning then possible. Solitary, conscious directed thought (on this standard picture) helps explain human rationality. It is at least part of the story of why humans know so much more than other animals; part of the story about why we can plan for the future, anticipate needs and problems, in the way other animals cannot. We can reflect on what we perceive and remember in order to leverage our informational resources. We can identify the subtle implications of what we already know. Conscious directed thought leads to better judgements and decisions.

On Mercier’s and Sperber’s view, the well-known dual systems view of cognition is a contemporary version of this standard story. On that picture, we have two modes of thinking. System 1 thinking is fast, automatic (requiring little top down attention and control) and tacit. We are aware of the result of a system 1 process, its output: the person on the left looks more trustworthy. But we are not aware of the intermediate steps in the process (the fact that their facial features are more regular). System 1 processes encode heuristics: reasonably reliable rules of thumb that can be rapidly applied. The heuristic nature of this reasoning explains many of the well-known foibles of human reasoning; for example, the fact that in general people struggle with the logically trivial Wasson selection task. System 2 reasoning is slower. It is demanding, so it is much more difficult to multi-task while engaged in system 2 thinking. And it is conscious, or mostly conscious. The agent is aware of the considerations on which her judgment is based as, say, she reads two CVs and decides which provides the better evidence of appointment to a position of trust. Non-human animals are exclusively (or perhaps almost exclusively) system 1 thinkers, and that explains the cognitive chasm that divides humans from their fellow animals.

Mercier and Sperber think this picture is completely wrong: the order of explanation is the wrong way around. As they see it, the primary function of thinking of reasons and about reasons is to present and assess reasons in public (and hence reasoning depends on language). The cognitive ability to consciously identify reasons evolved to allow us to persuade, coordinate and commit; to assess the proposals of others, thus reaping the rich benefits of cultural learning without being too vulnerable to exploitation; and to solve problems collectively when engaged in cooperative discussion. They base this view on the following considerations.

1. Mercier and Sperber argue that if the capacity to reason evolved to increase the reliability of judgements and of decisions through solitary, conscious deliberation, the design team should be sacked. If reasoning aims at truth-tracking and prudent decision making, it is
riddled with all manner of biases and inaccuracies. They make this case through a parade of cases — Linus Pauling on the merits of vitamin C, Bertillon on the guilt of Dreyfus and many others; a parade showing a profound insensitivity to objective considerations of evidence by unmistakably cognitively able agents. Moreover, these agents are engaging in slow, conscious, deliberative reasoning. Their biases cannot be explained by the short-cuts forced on system 1 reasoning. They rely as well on a wide-range of experimental results despite their rather sceptical take on this literature. They argue quite persuasively that there are serious questions about the ecological validity of many experiments, and that much of the experimental literature is excessively shaped by a syllogistic conception of appropriate reasoning. Even so, there seems persuasive evidence of systematic biases of various kinds; perhaps most obviously, humans are not natural Popperians, looking by default for disconfirmers of their initial views.

2. In contrast, the public use of reason is adaptive: we reason well (that is, in ways that would tend to track truth and prudence) in many collective contexts and when we assess the reasons of others. It is true that when we give reasons in justification for our views and actions, those reasons are much less reliably sound considerations in favour of the views advanced. But the role of reason-giving when presenting one’s own opinions is typically persuasive or justificatory (to retain or advance social credit) rather than truth tracking. So even when the justifications offered are tendentious, the giving of those reasons is often still adaptive. There are constraints: the reasons offered cannot be obviously absurd; cannot be obviously inconsistent with what has previously been asserted; they must be of a kind recognised in the community as potentially appropriate. “I was hungry” in our culture is not an appropriate reason for eating a dead pet. But these constraints leave open many reason-giving options that are not evidentially ideal. That said, it is probably easier to assess someone else’s case than develop your own view, and that might partially explain why it is easier to spot flaws in others’ opinions than to remove the flaws from one’s own. But that cannot be the whole story, for Mercier and Sperber have some intriguing experimental data of their own to show that we assess our own reasons better, when we can be induced to think of them as others.

3. Finally, Mercier and Sperber are deeply sceptical about the very idea of conscious deliberation: of the idea that we make judgements and come to decisions through inferential processes that we are aware of at the time, or can become retrospectively aware of, when prompted with “why did you think that?” On their view, retrospective explanation of our reasoning is always reconstruction, almost always incomplete, and very often with significant elements of confabulation. This is not scepticism about inference: inferential processes drive judgements and actions. But we are typically only aware of the outcome of our inferences, which often present as an intuition. Clearly, if solitary conscious deliberation is rare or non-existent, the public practices of giving, sharing and assessing reasons can hardly be parasitic on the private practice of conscious deliberation.

I am persuaded that when an agent gives reasons after a prompt or challenge, that is typically a highly fallible and partial reconstruction of the actual inferential pathway. It is not clear whether (or to what extent) this scepticism is meant to extend to occurrent deliberation: whether they deny that when we seem to be consciously working an issue through, we are veridically aware of the stages in the inferential process. If we do sometimes consciously deliberate, surely we could report, shortly thereafter, on that deliberation (especially if we have been using external aids like notes or diagrams)? An outright denial of the existence of

1 Perhaps even always: it is not clear exactly how strong the claim is.
solitary, conscious deliberation, of conscious directed problem solving, would be very implausible indeed. Surely agents do sometimes engage in explicit individual problem solving: working out whether (say) they need to stop for petrol by checking the fuel gauge, estimating the distance to travel, estimating their vehicle’s fuel efficiency. Mercier and Sperber could (and perhaps do) agree with this, but could then argue that these individual, internal cases are causally secondary to public reasoning. We evolved capacities for collective problem solving which can then be exapted for individual use. That idea is quite plausible in the light of the fact that the most persuasive individual cases involve public representational media; working out, say, how to get from one tube station to another by tracing a pathway with your finger on the map of the tube, while mentioning the relevant station names to yourself. Individual, explicit problem solving may be important and adaptive, but essentially involve recycling cognitive and social tools that evolved for use in the public domain.

Their core positive claim is very plausible. Mercier and Sperber are right to think that the practice of giving, assessing and discussing reasons has a natural adaptive fit to the unique conditions of human social life: with the importance of reputation (and hence consistency and defensible judgment); with the great potential benefits of information sharing so long as its risks can be managed; with the importance of coordination in collective and entangled action, and hence the need to negotiate plans and options. Agents with a social life like ours need something like the public practice of reason. Moreover, they are right to deny that public reasoning is just making public iterations of a cognitive process — conscious, solitary, explicit deliberation — that already existed prior to human social life, serving other functions. As I noted above, conscious explicit deliberation often, perhaps always, relies on public representational media (most obviously language), cognitive tools that only exist as the result of a distinctively human social life. Moreover, public reasoning does indeed seem much more powerful and efficient than the solitary version.

That said, it is a pity that they make no attempt to locate the evolutionary claim in a paleoanthropological context, nor develop any kind of incremental model of the emergence of public reason via some form of proto-reasoning from earlier hominins with fewer communicative tools. This is especially surprising, as their views would seem to complement the social brain hypothesis, defended by Dunbar and others (see for example Gamble, Dunbar et al. 2014)). Instead, they support their hypothesis by arguing that it predicts the existence of features of public discourse that we do indeed find. They then supplement that line of argument with evidence about the cultural universality and early onset of public reasoning. But while this evidence does provide defeasible support for the claim that public reasoning is an evolved feature of human life, it is neutral on the specific function that Mercier and Sperber isolate. Moreover, the fit between prediction and discovery is rather coarse. In particular, I am not persuaded that that the justificatory role of reasoning — reasoning to vindicate an opinion in response to challenge, and reasoning to persuade a sceptical audience — predicts a “myside bias”. They suggest that if public reason has evolved in part to allow an agent to seem reasonable in their views we should expect to find what we do indeed find. Agents much more readily think of considerations that support their previously stated conclusions than considerations that undermine that conclusion. I am sceptical. To the extent that agents are well adapted to ‘act as their own lawyer’, they should

---

2 Only defeasible support: Celia Heyes argues that an array of widely distributed and early developing cognitive capacities are acquired through social learning, and are the result of cumulative cultural evolution, not genetic evolution. See her (Heyes 2018)
be alert to potential problems, so they can defuse them. I do not see this as a deep challenge to the core proposal. For I suspect public reason is something of a kludge. It depends on more or less the full resources of contemporary language, and while it’s a response to a social world that had slowly become more complex, the most profound changes are probably relatively recent. Public reason may have become a systemic and important part of human social worlds over the last 100k years or so, and used in a varying mix of competitive, partially competitive, and fully cooperative interactions (Sterelny 2011, Sterelny 2014). If that is right, the capacities that support public reasoning evolved recently, and under somewhat conflicting selection pressures. For a version of public reasoning capacities optimal for cooperative interaction would not be optimal for competitive ones. Assessing reliability (is the interlocutor likely to be an expert on X?) calls for quite different probes than assessing manipulative intent (is the interlocutor’s interests likely to be in sharp conflict with mine?). We can expect design compromises, exacerbated by building the relevant machinery from whatever previous evolutionary history had made available. Given all this, we would expect a coarse fit between function and performance. Thus I think the positive case for their function claim is generally persuasive. That said, they overstate the case against the importance of private, individual reasoning. I also think that their plausible evolutionary claim is embedded in an implausible model of the cognitive system.

2. The Intellectualist View of Reason

As Mercier and Sperber see it, the standard view of reason equates the human capacity for reasoning (in the sense of conscious deliberation) with human rationality. This in turn is the explanation of the cognitive divide that distinguishes us from other animals, and which explains our success as a species. In my view, they overcook their scepticism about this in three ways.

First. They see it as a hypothesis about the mind working in splendid isolation (hence their final debunking chapter about solitary genius). But while solitary, conscious deliberation very probably did not precede and explain deliberation in the public domain, there may well have been positive coevolution between the two. Individuals’ capacities to solve their own problems by conscious deliberation may have improved, and been improved by, participation in public deliberation. Indeed, to the extent that agents think (consciously) before they talk or write, the two forms of deliberation are closely intermingled.

Second. They saddle the model with unmotivated commitments. They treat the intellectualist model as committed to the idea that deliberation harnesses a deduction engine in which every premise and intermediate step of the deduction is explicit, but so too are the inferential

---

3 Mercier and Sperber discuss this response, and argue that it is much less cognitively demanding to respond to objections than to anticipate them, as objections might well depend on particular features of the audiences’ preferences, or on information that only they are privy too. I am still not persuaded. First an anticipated objection is likely to be answered much more convincingly (forewarned is forearmed). Second, while it is sometimes true that objections depend on private information, they also often flow from obvious interests and from public knowledge.

4 It is noteworthy that they cite very few contemporary defenders of this standard view. As Anton Killin has pointed out to me, it is somewhat surprising that they have not cited Chomsky, with his view of the primacy of i-language over e-language, as a foil for their view; for example (Berwick and Chomsky 2016).

5 Perhaps supplemented with probability theory.
principles which connect premises, intermediate conclusions, and conclusion. Consider, for example, the following passage (p131):

As we pointed out in Chapter 7, it takes more to have a reason than to just recognize some fact. You can walk out and see that the pavement is wet, but you cannot just see that this is a reason to believe that it has been raining. That the pavement is wet may be an objective reason for concluding that the pavement is slippery, that the outside temperature is not below freezing point, that one’s shoes will get dirty, and so on. To think, “The pavement is wet” is not by itself to entertain a reason. You may, moreover, intuitively infer that it has been raining from the fact that the pavement is wet without this relationship between premise and conclusion being mentally represented in the process. Only if you were to entertain a thought like “From the fact that the pavement is wet it follows that it must have been raining” would you be recognizing the reason for your conclusion.

This is strange. As they know, no such deduction engine can be built, with all the inferential principles appearing as extra premises, on pain of infinite regress. So they seem to saddle the intellectualist view with an incoherent model. The friend of solitary conscious deliberation is committed to the claims that (a) we sometimes engage in conscious, directed inference (or problem solving) in which the agent is consciously aware of the premises on which he/she relies, and of the steps of the problem solution. But the agent need not be aware of the inferential principles in use, nor of the search mechanisms through which premises are bought into working memory. Still less need they represent those principles as premises. And (b): this sometimes helps. We sometimes solve problems through such deliberation that we could solve no other way. Mercier and Sperber could, and I think should, accept (a) and (b): all they need deny is that this capacity evolved prior to, and explains, public reasoning.

Third, there is a curious tension in their own view. For it is part of their overall picture that the mind is massively modular. We do not genuinely have a capacity for domain general inference. But reasoning, as part of our metarepresentational capacity, helps give us a surrogate for domain general reasoning. For the mechanism through which we extract (and assess) reasons for the public space has multi-modular access; it takes input from other modules. But surely there is a deep tension between the claims that (a) metarepresentation, including reasoning, gives human agents a kind of virtual domain general reasoning, and (b) reasoning has neither the function nor the reliable effect of upgrading individual cognitive competence; reasoning does not help you think better. It is surely a very considerable upgrade (and one likely to have selective consequences) if the capacity to represent our own and other’s representations (and to make them public) confers on agents a capacity for domain general, or cross-modular, inference that they would otherwise lack.

That said, it is not at all clear to me to how metarepresentation is supposed to give domain general capacities on the cheap. Mercier and Sperber rightly make much of the fact that in the right circumstances we use others as instruments to find out about the world; metaphorically speaking, we see through their beliefs to the world. From the fact that I represent Hugo as believing that it is raining outside, I come to believe myself that it is raining outside. However, my ability to piggyback on Hugo’s beliefs presupposes that I understand them. If I lack a basic conceptual understanding of molecular biology, I cannot learn molecular biology by representing Hugo’s beliefs about the protein structure of collagen. If Hugo can integrate information between two domains (perhaps natural history and the social world), and I lack that capacity, I will not be able to represent Hugo’s beliefs about the importance of team
organisation in foraging tasks. Of course I can learn conceptually from Hugo as well as factually. But unless he is engaged in a project of systematic instruction, I’ll need to understand most of what Hugo says to understand the context that will help me fix on the bits I do not understand. The same issue arises with the reasoning module if it is thought to access and combine information from different modules in generating justification or assessment. If the representations in different domains are in different formats, and are adapted to different computational procedures, it is not at all obvious that a reasoning module can build an integrated representation from elements extracted from different modules; not unless the reasoning module has most of the capacity of a Fodorian central processor (or perhaps a set of special purpose translation schemes, on for each module, mapping each module’s output into a common format). In short, the way in which metarepresentational capacities lead to virtual domain generality remain unclear to me. But if they did, surely this would be a major advance in individual cognitive capacity?

3. Strictures on Modularity

Mercier and Sperber give a crisp account of reasoning as a public phenomenon. They could have developed this account of the proper function(s) of reason without committing themselves to a view of the cognitive mechanics which make public reason possible. Likewise, their scepticism about the nature and reliability of our introspective access to the causal sources of our opinions does not depend on a positive account of retrospective or occurrent conscious deliberation. However, without quite committing themselves to a cognitive model of the internal sources of public reasoning, they go close to endorsing a modular account of reason. Public reasoning depends on a metarepresentational module that performs (non-conscious) inferences on representations sourced from other modules, producing an intuitive judgement as an output. This specific hypothesis is framed through their endorsement of a modular model of cognitive organisation in general. I am sceptical both of the overall framework and the specific hypothesis. I begin with the framework.

What Does a Modular Conception of the Mind Explain? In introducing their modular approach to the mind, Mercier and Sperber quite reasonably point out that they are not committed to a Fodorian conception of modules: encapsulated, cognitively impenetrable, largely innate, autonomous, disassociable from other cognitive functions, probably somewhat neutrally localised. That is true, but the Fodorian conception is not arbitrary. The idea (i) that modules are encapsulated and (ii) that representational vehicles and the procedures that access them are tuned in advance to a particular domain, and to solving specific questions about that domain, are supposed to jointly make plausible the idea that modular operations are computationally tractable (Fodor 1983). Modular evolutionary psychology took over a roughly similar idea. Specialist devices outperform generalist devices because specialist devices are pre-wired to access the right kind of data, represent it in the right format, and probe it with the right questions (Tooby and Cosmides 1992).

Mercier and Sperber do not need to buy into any of this. But in downplaying encapsulation and pre-wiring, their modules are as computationally opaque as a general purpose problem solver about which they are so sceptical. For example, in talking about theory of mind modules, they note (p97-98):

---

An encapsulated device has access only to a sharply limited subset of all the information available to an agent; a cognitively impenetrable device is insensitive to top down cognitive control.
“We humans tend to constantly monitor our social environment (as in the waiting room example). We open, maintain, and update “mental files” on all the people we know (including people we only know of, like kings or famous actors, not to mention fictional characters whose thought we also know quite a bit about). In these files about people (and other kinds of agents such as caterpillars and gods), there may be all kinds of information: information about their names, family, history, appearance, dispositions, doings; information also about what is in their mental files where they store information about other people, us included. Our mental files about other people contain information about the contents of their mental files, and that information is provided by mindreading (and is, of course, metarepresentational).

Some of your mental files about people are very thin and short-lived, such as the file you opened about the other patient in the doctor’s waiting room. Other files are thick and permanent, such as files about members of your close family. Some of the mindreading information in these files is provided by your spontaneous interpretation of what you observe others doing. Some information is provided by the people themselves who, in communicating, help you read their minds. Further information is provided through people talking about people: gossip. The point is that we read minds on the basis of a great variety of evidence and in a great variety of ways.”

Some module! It follows that the Sperberian hypothesis about cognitive architecture (taken over from (Sperber 1996) does not offer a prima facie solution to frame-type problems. His architecture does not divide a supposedly unified, domain general, but computationally mysterious capacity into more limited, specialised but less computationally mysterious components. Modules as they appear here are still computationally mysterious.

I take their positive case for modularity to be the argument that our inferential engines rely on domain specific regularities about the world that they do not represent. For as I read The Enigma of Reason, the crucial argument for modularity appears around pp 85-87. It comes after a discussion of early infant understanding of simple material phenomena as an automated folk physics module (one module I do find plausible). Our inferential capacities are parasitic on regularities in the world, regularities that vary in stability and generality. When there is a salient regularity $X \rightarrow Y$, we become tuned to infer $Y$ on the recognition of $X$. But we do not have to (and could not in every case) represent that regularity and use it as a premise in a deductive argument. This is clearly right. It is very likely that we would not represent the regularity explicitly if that expectation is hard-wired, and the same is probably true of regularities to which we become sensitised by associative learning (see p 122). Often our inferential habits depend on these unrepresented regularities. Since these regularities are features of natural domains (like the physical interactions of solid objects), our various inferential systems (the argument runs) will be tuned by evolution and learning to specific domains. If the regularities on which our inferences relied were represented as premises (as they sometimes are), we could have a single set of logical and probabilistic inferential strategies, taking input from many natural domains as data, and spitting out conclusions. But representing a regularity and using it as a premise in an inference is probably the exception, not the rule. Thus according to this line of thought, the more inference is implicit, and the less the causal structures of human environments are explicitly represented, the more modular and domain specific our inference system must be.

I am persuaded that inference must often depend on regularities that are not explicitly represented as premises. But I do not think it follows that we have a set of inferential systems
rather than a single integrated system. Fido does not have two inferential modules, one to handle food, and the other to handle social interaction, because Fido can associate the sound of the key in the door with his owner returning, and associate the sound of the fridge door opening with food, without representing either causal regularity. I think that Mercier and Sperber assume that the defenders of a domain general inference system are also committed to a formal, logicist conception of how inference works: that the inferential connections between representations depend solely on the formal, abstract structure of those representations; on their “logical form”. But that view is not compulsory, and for that reason we may have a single integrated system in which information is represented and is available for inference, yet a system in which inferences can exploit local regularities. There is a tradition that dates back at least to Carnap on meaning postulates; one that has insisted that some entailments depend not on abstract structure but on the meaning of specific lexical items. Such postulates give us local predicate-specific inferential rules like “if X is red” → “X is coloured”. (Carnap 1952). Local causal contingences can shape inference without being represented as explicit premises. A domain general cognitive architecture might instead have a single but rich set of inferential strategies; strategies which include local, meaning-postulate-like rules that tune the agent to reasonably stable, reasonably general local regularities, as well as inferential rules that are sensitive only to the general formal structure of a representation. Indeed, Mercier and Sperber themselves give plenty of evidence that human inference is not well captured by formal models in which validity and invalidity depend only on the abstract logical forms of premises and conclusion, with specific predicts and referring terms being irrelevant to logical connection. On the alternative suggested here, inference is an integrated, single system, but with the output of each local rule, and the many more general ones, being a potential input to the others.

A Metarepresentational Module? So I remain sceptical about the view that the mind is largely modular in organisation. But even setting that aside, the idea that public reasoning depends on a metarepresentational module never comes into clear and plausible focus. First, Mercier and Sperber seem to have a non-standard view of metarepresentation. Their claim is not explicit, but they seem to think that a module is metarepresentational if the inference depends on features of the representational vehicle. This comes out in their discussion of an arithmetical example, where they point out that it is trivially easy to infer that 900 is the product of 300 x 3, but the triviality of this inference depends on our representing these numbers in base 10 (see pp 101/102). They suggest as a consequence that arithmetical inference is metarepresentational; for it is sensitive to the format of the representation, not just its informational content. But though this arithmetical inference depends on properties of the way we represent quantities, it does not follow that arithmetical inferences are about the representation. They are about the quantities themselves; a distinction that Mercier and Sperber are careful to draw themselves on the next page. Think not of quantity but of space. The use of maps makes some inferences about distance and relative position trivial. But they are still inferences about (say) the relative locations of tube stations, not about the tube map. On any computational theory of mind, all inferential processes are sensitive to the format of representations, for they are keyed to the structural features of representations (Fodor 1975). That does not make all computational modules metarepresentational modules.

---

7 They do not say this explicitly. But their discussion of this example is immediately preceded by the remark that “Numerical cognition, for instance, provides us with a sharply different kind of metarepresentational intuitions.” (p101)
A similar puzzle arises with the supposedly metarepresentational reason module. Consider reason under challenge. I am asked why I chose a particular café and I respond by saying “it is close and cheap”. My reason — that the café is close and cheap — is about the café, it is not about the representation of a café. So when I give a reason, I am not producing a metarepresentation. No doubt there was some tacit search process, and that search process was sensitive to the features of my various representations of the café. But just as in the arithmetic case, the sensitivity of the inferential processes to features of the representational system does not make the output of that process about the representation. Of course evaluating reasons clearly is metarepresentation. When an agent says, or thinks, that a reason is implausible, they are representing the reason itself. That might be true of giving reasons too, if that was the result of conscious deliberation in which the agent evaluates and assesses candidate reasons before offering one. But that is not Mercier’s and Sperber’s picture of what typically happens when we offer reasons. We confabulate when we offer such a picture in justification of our opinions. On their view, the production of reasons in response to a query is a tacit process, opaque to introspection. Mercier and Sperber could say that assessing reasons is genuinely metarepresentational but offering reasons typically is not. Moreover, that would suggest an explanation of why (from the point of view of epistemic reliability) we are so much better at assessment than justification. But that would give up the view that the production of reasons and their assessment depend on the same cognitive mechanism.

Perhaps giving up a unified view of production and assessment is not much of a cost, but even if one gives up a unified picture, it is difficult to see how the metarepresentational module could work. Mercier and Sperber draw many of their examples of tacit inference from perception. For example, on pp 58-59, they discuss a lovely example of Roger Shepard of an apparently larger monster chasing a smaller one. In fact, the monsters on the 2D image are of the same size, but the context induces us to see it as a three dimensional scene, with one behind the other, and so further away. On that 3D reading, our cognitive system rapidly and seamlessly infers that because the more distant figure projects the same size retinal image as the closer one, it is actually larger. So we see it that way. Our inference exploits the very dependable physical principle that relates relative size to distance, and also to the cues that induce us to see a flat image three dimensionally. It does so of course without representing that physical principle and using it as a premise. Mercier and Sperber suggest, likewise, that the reasoning module relies on unrepresented empirical generalisations in tacit inferences that connect representations to reasons (or publically presented reasons to assessment of reasons). But what kind of empirical regularities are there that relate representations, especially when they can be sourced from different modules? In summarising their proposal, they say: "Part of the answer is that our first-order intuitions (about Molly’s mood, the rain, and the vast variety of things about which we have such intuitions) are delivered by a great many modules, while our metarepresentational intuitions about reasons for our first-order intuitions are delivered by one metarepresentational module that just works on reasons. First-order modules draw all kinds of inferences about objects in their domain of competency by exploiting regularities in their domain. The metarepresentational module involved draws inferences in its domain of competence; it draws, that is, inferences about the relationship between reasons and conclusions. To do so, it attends to relevant properties of this relationship.” (p 145)

But what are those relevant properties? They do not say, but attending just to the final part of the summary, one might think they are logical and quasi-logical properties (like relevance). Those are properties of the relationship between reasons and conclusions. But while that fits
the supposed role of the reasoning module is assessing reasons, it does not fit the giving of reasons mentioned in the first part of the quote about first order intuitions. In this case, the intuition that Molly is upset. In such cases, the metarepresentational module is not recovering the history of the formation of that intuition. Mercier and Sperber vehemently reject the view that justification is retrospective access to the actual inferential process that produces a judgement. Rather, the module constructs a vindication of that intuition. Yet until it does such a construction, there is no relationship between reasons and first order intuition to have properties, relevant or not. Moreover, it is hard to see how the logical properties of representations could play a central role, for logic is permissive. Whatever a reason module might do, it is not noting all the logical implications amongst the representations it can access, to determine whether one or more entail the opinion for which it is seeking vindication. The process must be directed by some form of search, and that search looks computationally mysterious.

This computational problem is exacerbated if the various modules that contribute to human thinking by delivering first-order intuitions are sharply distinct from one another; that is, if the representational vehicles in the different modules are optimised for the extraction and use of particular kinds of information. Contrast maps, ideal for much spatial reasoning, with pie charts, good for representing quantitative and proportionality information. This example illustrates how different the inputs to a reasoning module might be. Just because they are all, in some sense, representations, it does not follow that they are in a common format, or that they can be processed in the same way. Perhaps the thought is that the reasoning module has access only to the outputs of the module, and these must be in a common enough format because they can be expressed as public opinions, But I can see and report on both maps and pie charts. That does not imply that they have a common format. Of course everyone has the problem of explaining how we think of the reasons that supposedly justify our intuitive opinions. That said, the appeal to a metarepresentational reasoning module seems not to help at all in solving this problem. Indeed, because there is no equivalent of a central processor where information from the modules is bought together in a common format, it might even make it worse.

In brief. I am very much on board with the evolutionary hypothesis, but not with the cognitive science framework.8

References

8 Thanks to Celia Heyes, Anton Killin, Stephen Mann, and Ron Planer for comments on an earlier draft of this essay, and to Hugo Mercier for e-mail correspondence on it. Thanks also to the Australian Research Council for their generous grants in support of my research on the evolution of human cognition and social life.