NB: The following paper present the work at the basis of our presentation to the Collective Wisdom conference, and for our contribution to the proceedings, but is not, in its present form, this contribution

(draft – not to be quoted)

Hugo Mercier & Dan Sperber

Intuitive and reflective inferences

Introduction

Experimental evidence on reasoning and decision making has been used to argue both that human rationality is adequate and that it is defective. The idea that reasoning involves not one but two mental systems (see Evans & Over, 1996; Sloman, 1996; Stanovich, 2004 for reasoning, and Kahneman & Frederick, 2005 for decision making) makes better sense of this evidence. 'System 1' reasoning is fast, automatic, and mostly unconscious; it relies on 'fast and frugal' heuristics (to use Gigerenzer's expression, Gigerenzer, Todd, & ABC Research Group, 1999) offering seemingly effortless conclusions that are generally appropriate in most settings, but may be faulty, for instance in experimental situations devised to test the limits of human reasoning abilities. 'System 2' reasoning is slow, consciously controlled and effortful, but makes it possible to follow normative rules and to overcome the shortcomings of system 1 (Evans & Over, 1996).

The occurrence of both sound and unsound inferences in reasoning experiments and more generally in everyday human thinking can be explained by the roles played by these two kinds of processes. Depending on the problem, the context, and the person (the ability for system 2 reasoning is usually seen as varying widely between individuals, see Stanovich & West (2000)) either system 1 or system 2 reasoning is more likely to be activated, with different consequences for people's ability to reach the normatively correct solution (Evans, 2006). The two systems can even compete: system 1 suggests an intuitively appealing response while system 2 tries to inhibit this response and to impose its own norm-guided one.

Much evidence has accumulated in favour of such a dual view of reasoning (Evans, 2003, in press; for arguments against, see Osman, 2004). There is, however, some vagueness in the way the two systems are characterized. Instead of a principled distinction, we are presented with a bundle of contrasting features—slow/fast, automatic/controlled, explicit/implicit, associationist/rule based, modular/central—which, depending on the specific dual process theory, are attributed more or less exclusively to one of the two systems. As Evans states in a recent review, "it would then be helpful to have some clear basis for this distinction"; he also suggests that "we might be better off talking about type 1 and type 2 processes" rather than systems (Evans, in press).

We share the intuitions that drove the development of dual system theories. Our goal here is to propose in the same spirit a principled distinction between two types of inferences: 'intuitive inference' and 'reflective inference' (or reasoning proper). We ground this distinction in a massively modular view of the human mind where metarepresentational modules play an important role in explaining the peculiarities of human psychological evolution. We defend the hypothesis that the main function of reflective inference is to produce and evaluate arguments occurring in interpersonal communication (rather than to help individual ratiocination). This function, we claim, helps explain important aspects of reasoning. We review some of the existing evidence and argue that it gives support to this approach.

Inferential processes and massive modularity

Dual process theories stand in contrast to more traditional monistic views that assume that reasoning is governed by a single system, be it one of rules (Braine, 1990; Rips, 1994), or mental models (Johnson-Laird, 1983). At first blush, dual process theories also stand in contrast to massively modular views of human cognition (Barrett & Kurzban, 2006; Carruthers, 2006; Sperber, 1994; Tooby & Cosmides, 1992). Massive modularists are neither monists nor dualists, they are pluralists. They see the human mind as made up of many specialised modules, each autonomous, each with a distinct phylogenetic and/or ontogenetic history, and each with its own input conditions, specific procedures, and characteristic outputs.

In the human case, most innate modules are learning modules (e.g. the language faculty) and perform their function by using environmental inputs to construct acquired modules (e.g.

the grammar of a particular language). Given the prevalence of these innate learning modules, massive modularity does not imply massive innateness of mental modules: many or most of them are the output of an acquisition process. This is not the place to argue in detail for the massive modularity thesis (but see Sperber, 1994, 2001b, 2005). What we want to do rather is to explore some implications of the thesis for the psychology of reasoning, and in particular for the interpretation of the kind of phenomena that has inspired dual process theories.

Massive modularists assume that inferences are carried out not by one or two systems but by many domain-specific modules that take advantage of the peculiar regularities of their specific domains to apply inferential procedures that would be inappropriate in other domains.

If, for instance, we had seen two objects being put behind an opaque screen, we are surprised to see only one object when the screen is lifted (and so are 12-month-old infants, Wynn, 1992). We expected there would be at least two objects. An expectation is the outcome of an inference. In drawing this inference (in a typically unconscious manner) we do not use as a premise the assumption that solid objects persist through time. We passively ignore rather than actively deny the possibility of their vanishing or blending with one another. This assumption of persistence is built into a domain-specific mechanism we use to draw inferences about solid objects. No such assumption is built into the way we draw inferences about, say, liquids. If we see two liquids being poured in the same opaque vessel, we are not surprised to discover only one liquid when the content of the vessel is made visible.

Similarly we have different built-in assumptions about the fall of objects and their other changes of location (Spelke, 1990), about the movement of animate and inanimate objects respectively (Leslie, 1995), and about the relevance of information unintentionally made available and that of information intentionally communicated (Sperber & Wilson, 2002).

Massive modularity may seem incompatible with the sense we have that our thinking is a unitary and integrated process. However, this vague introspective datum is of no serious evidential value. More relevant is the fact that we can and do reason on premises that pertain to more than one cognitive domain. For instance, when we see a small child (but not a cat) sitting on window sill (but not on a low bench), we infer that there is a serious and pressing danger, unproblematically integrating premises from our knowledge of child psychology and form our commonsense knowledge of physics. A sensible massive modularity theory must however assume that the same premise can be processed successively or in parallel by several modules, just as the same food can be decomposed successively or in parallel by several enzymes (for a development of the analogy between enzymes and modules suggested in

Sperber 1994, see Barrett, 2005) Multi-domain inferences can be the joint work of several domain-specific modules.

Fodor (2001) has developed a more serious objection to the idea that human inferences could be performed by a massively modular mind. Individual modules, by their very nature, have very little or no context-sensitivity. Human inference on the contrary is characterized by high context-sensitivity: the same input can yield quite different conclusions in different contexts. Another way to put the same point is that human inference tends to process not just any available inputs but only the most relevant ones in the situation, and tends moreover to contextualise each of these inputs in a way that maximises its relevance. This is arguably a major feature of human cognition (described in Sperber & Wilson, 1995 as the "cognitive principle of relevance"). If one assumes, as does Fodor, that the operations of modules are mandatory, in the sense that they automatically process any input that meets their input condition, then human cognition should be stimulus-driven, with the same stimuli triggering the same inferences in all contexts, and the output of these inferences triggering the same higher level inferences in all contexts, and so forth. The high context-sensitivity actually exhibited by human inferences provides a powerful argument against the view that the human mind is massively modular in the way envisaged and criticised by Fodor.

In fact, however, it is dubious that any cognitive processes can be considered mandatory in the intended sense. As studies on attentional blindness demonstrate, even when the psychophysical conditions for perception are fully met, some outstanding stimuli may remain unperceived (e.g. a person disguised as a gorilla moving in full view in the middle of a few basket-ball players; see Simons & Chabris, 1999). Thus even perceptual mechanisms—which for Fodor are prototypical modules—do not automatically process every input that meets their input conditions. This is easily explained. Human cognition is characterised by the fact that, at any moment, it is monitoring the environment and has available in memory much more information than it could simultaneously process. 'Attention' refers to the dynamic selection of some of the available information from the environment and from memory for deeper processing.

From a modularist point of view, attentional selection might be best seen, not as the output of a distinct attention mechanism allocating resources to specific modules, but as the result of a process of competition for such resources among modules. Some modules, for instance danger detectors, may be permanently advantaged in this competition because their inputs have a high expected relevance. Other modules may be advantaged at a given time because of a decision to attend to their potential inputs. For instance, face recognition is on

the alert when waiting for a friend at the train station. Leaving aside these permanent bottom-up biases and temporary top-down biases, modules with the highest level of immediate activation both from upstream and downstream modules should be winners in the competition (with ongoing changes in these levels of activation resulting in shifts of attention).

A competitive system of this type fine-tuned both in phylogenetic evolution and in individual development would go a long way towards explaining how human cognition can, in practice, tend toward high context-sensitivity or, equivalently, towards the maximisation of the relevance of the inputs it processes.

Still, there seem to be some inferential processes that are truly domain-general. For example, people seem to be able to infer *Q* from *P-or-Q* and *not-P* whatever the content of *P* and *Q*, be it concrete or abstract, factual or imaginary. Similarly, people are able to infer from *X believed that P* and *X now believes that not-P* that X has changed her mind regarding the subject-matter of *P*, whatever this subject-matter. Or again, people are capable of inferring what a speaker means from what she utters, whatever she is talking about. Are these genuine instances of domain-generality (and therefore of non-modularity)? A more careful examination of the inferential mechanisms involved reveals that they are as domain-specific as any other cognitive mechanisms; they just happen to draw inferences warranted by the properties of a very peculiar kind of objects: conceptual representations.

There is a standard distinction between *perceptual* mechanisms that have as input sensory data and as output representations of distal stimuli, and *conceptual* mechanisms that have representations both as input and as output. Still, just as perceptual mechanisms are not drawing inferences about sensory data but about perceived objects, conceptual mechanisms are not drawing inferences about the properties of the representations they process but about the properties of the objects or states of affairs represented in these representations. For instance, inferring from the perceived presence of dark clouds that it will rain is not an inference about that perception but about the clouds themselves and their likely effects.

Representations, be they mental (e.g. beliefs) or public (e.g. utterances), are also objects in the world. They have properties qua representations. The belief that it will rain is a mental representation held by John at a given time and given up by him at another; it may be consistent or inconsistent with some of his other beliefs; it may be true or false; and so on. Consistency and truth are properties not of the state of affairs represented but of the belief itself. Mary's utterance is a public representation that may be grammatical or not, relevant or not, addressed to John or to Jean. These are properties not of the state of affairs talked about but of the utterance itself. The inference of a conclusion of the form *Q* from a pair of premises

of the form *P-or-Q* and *not-P* is warranted not by the state of affairs described by these premises but by the formal properties of these representations considered in the abstract.

Humans are aware of the existence of representations in the world, and a good part of their behaviour is aimed either at influencing the mental representations of others or at improving their own mental representations. From a modularist point of view, it is sensible to expect that the special properties of representations (or of specific types of representations) should be exploited by modules that specialise in drawing inferences *about* representations. For this, representations have themselves to be represented by means of second-order representations, or metarepresentations (see Sperber, 2000b).

Modules that draw inferences about representations are metarepresentational modules. The representations about which metarepresentational modules draw inferences can themselves be about any subject-matter within or across any cognitive domains. Inferences about representations on a given subject-matter may often provide reasons to accept specific conclusions about that subject-matter. For instance, knowing that a competent meteorologist believes that it will rain is a reason to believe that it will rain. Or knowing that a set of premises entails a conclusion is a reason to believe that conclusion if one believes the premises.

The fact that metarepresentational inferences may indirectly yield conclusions that belong to the domains of the representations metarepresented results in a semblance of domain-generality. However, since metarepresentational mechanisms only process specific properties (e.g. who is entertaining a given representation or what a set of representations entails) of a specific kind of objects (representations), this is only an *indirect and virtual* domain-generality. Metarepresentational modules are as specialized and modular as any other kind of module. It is just that the domain-specific inferences they perform may result in the fixation of beliefs in any domain. In this respect, metarepresentational inferences are comparable to visual or auditory perception. Visual perception mechanisms are highly specialised and attend to special properties of highly specific optical inputs and yet they may fixate beliefs in most cognitive domains.

Massive modularity is clearly incompatible with a monistic view of inference: if inferential procedures are carried out by many different modules using a variety of procedures, then it is pointless to ask how inference in general is performed or to try to generalise from the properties of inference in a given domain to all inferential process. For instance, it could be that spatial reasoning of some kind is performed by means of mental models but that, nevertheless, mental models play no role in most or even all other inferential

processes. Is massive modularity similarly incompatible with a dualistic view? We will argue that it is not and that, in fact, a modularist approach provides a principled way to develop such a view.

Intuitive and reflective inference

Inferential modules alter—and if things go well, improve—the information available to an individual by adding new beliefs, updating or erasing old ones, or modifying the strength, or the subjective probability, of existing beliefs. These modifications occur at what Dennett called the 'subpersonal' level (Dennett, 1969); see also Frankish, this volme) They are the output of processes that take place inside individuals without being controlled by them. The modification of the stock of beliefs (or the 'data base') that results from spontaneous inference occurs without the individual's attending to what justifies this modification, just as in the case of perceptual processes.

As we pointed out, there are different kinds of representations. Some properties are shared by all representations, others are specific to one given kind. From a modularist point of view, it is sensible to ask whether the different inferential opportunities offered by various types of representations are taken advantage of by several distinct metarepresentational modules. Yet, in the literature, metarepresentational abilities are generally equated with a single 'Theory-of-Mind', 'mentalization', or 'mindreading' module. 'Metarepresentational' (i.e. about representations) is treated as more or less synonymous with 'metapsychological' (i.e. about mental representations). This is both too broad and too narrow.

It is too broad because some types of attributions of mental states are best performed by means of specialized inferential routines rather than by a unitary general mindreading ability. For instance the intention of another person to establish joint attention with you can be inferred from a fairly simple, possibly repeated, action sequence in which eye contact with you is followed by staring at the intended target of joint attention. 9-month-old infants are capable of using this behavioural pattern to attribute an intention of this kind. For this, they do not use a general mindreading ability but a much more specialised module with, presumably, a strong genetic basis (Baron-Cohen, 1995; Tomasello, 1999).

Treating 'metarepresentational' as synonymous with 'metapsychological' is also too narrow since metarepresentations are used to represent not just mental representations but also public representations, such as utterances (Wilson, 2000), and representations considered in

the abstract, independently of their mental or public instantiations, as in logical or mathematical reasoning.

There are good evolutionary reasons (discussed in the next section) to make the assumption that, among metarepresentational modules, there is one specialised in argumentative relationships among conceptual representations (Sperber, 2000a, 2001a). Often, we are interested not just in some claim (for instance the claim that it will rain this afternoon) but also in reasons to accept it (for instance the fact that there are heavy clouds) or to reject it (for instance the fact that last weather bulletin forecasted clouds but not rain) . This occurs in two types of situations: somebody is making a claim that would be relevant to us if it were true but we are not disposed to accept it just on trust, and so we look at reasons to accept or reject it; or we are trying to convince an interlocutor of a claim that she won't accept just on trust, and therefore we have to give her reasons to accept it.

What the argumentation module does then is to take as input a claim and, possibly, information relevant to its evaluation, and to produce as output reasons to accept or reject that claim. The workings of this module are just as opaque as those of any other module, and its immediate outputs are just as intuitively compelling. We accept as self-evident that a given pair of accepted assumptions of the form *P-or-Q* and *not-Q* justifies accepting the conclusion *P*, but this compelling intuition would be hard to justify.[1] We accept as self-evident that, everything else being equal, we are likely to be better off betting on a horse that has won many races than on a horse that has won few races, but as philosophers since Hume have argued at length, we would be hard put to justify this type of compelling intuition (see (Vickers, 1988). Still, the argumentation module provides us with reasons to accept conclusions, even though we may be unable to articulate why we accept these reasons as reasons.

The direct output of all inferential modules, including the argumentation module, is intuitive in the clear sense that we just trust our own mental mechanisms and that we are disposed to treat as true their output without attending to reasons for this acceptance, or even without having access to such reasons.

In the case of the argumentation module, however, there is a subtle twist that, if not properly understood, may cause confusion. The intuitive output of the argumentation module

[1] Arguing that the inference is justified by the logical properties of the connectives 'or' and 'not' is not enough. Arthur Prior has imagined a connective, 'tonk', defined by two rules: (1) [P→(P tonk Q)]; (2) [(P tonk Q) → Q]. With 'tonk', one may then infer any proposition Q from any proposition P. This is of course inacceptable and illustrates the point that only appropriate connectives permit sound inferences. This in turn raises the difficult question of what makes a connective appropriate (see (Bonnay & Simmenauer, 2005; Engel, 2006; Prior, 1960)

consists in the representation of a relationship between a conclusion and reasons to accept it. This representation is produced by a communicator aiming at convincing her audience, and evaluated by her audience unwilling to be convinced without good grounds. Here, for reasons of space, we consider only the audience's perspective. For the audience, intuitively accepting the *direct* output of the argumentation module, that is the representation of an argument-conclusion relationship, provides explicit reasons to accept on its own the conclusion embedded in it. The acceptance of this embedded conclusion, when it occurs, is an *indirect* output of the argumentation module.

At the level of personal psychology, there is a major difference between intuitively accepting some representation as a fact, and accepting some claim because of explicit reasons. In the second case only, do we experience engaging in a mental act that results in a conscious decision to accept. At the level of subpersonal cognitive psychology, disembedding a conclusion from the argument that justifies it is not, properly speaking, an inferential operation—it does not result in a new conclusion—but a data-management one. It allows a conclusion that has already been derived to be stored and used on its own.[2] What, at a personal level, looks like a decision to accept a conclusion, is, we suggest, realised at the subpersonal level by this data-management operation. This indirect output of the argumentation module—the disembedded conclusion—is quite unlike the direct output of this and all of other inferential modules in that we mentally represent a reason to accept it. Conclusions accepted for a reason are not intuitive but are, we will say, 'reflective' (Sperber, 1997) and the mental act of accepting a reflective conclusion through an examination of the reasons one has to do so is an act of reflection.[3]

There is thus, within a massive-modularist framework, a subtle but unambiguous way to distinguish two categories of inferences: intuitive inferences the conclusion of which are the direct output of all inferential modules (including the argumentation module), and reflective inferences the conclusions of which are an indirect output embedded in the direct output the argumentation module. Since reflective inferences involve the representation of reasons, they well deserve the name of reasoning proper.

In this perspective, the sense we have that reasoning is a slow and effortful mental process does not come from the difficulty of the individual reflective steps involved, but from

---

[2] Dan Sperber (1985, chapter 2) has suggested that there are conditions of intelligibility on such disembeddings and that poorly understood conclusions are stored inside their validating context and not on their own.

[3] Several philosophers (Cohen, 1992; Engel, 2000; Stalnaker, 1984; see also de Sousa, 1971; Dennett, 1981) have proposed a contrast between belief and acceptance that is interestingly similar to ours between intuitive and reflective conclusions.

the fact that typical reasoning involves a series of such reflective steps. The conclusion embedded in an output of the argumentation module is disembedded and used as part of the input for another operation of the same module, and this can be reiterated many times. The difficulty of reasoning comes from the attentional or 'concentration' effort needed to maintain long enoughan expectation of relevance strong enough to keep the argumentation module active throughout this series of steps when other modules are competing for energetic resources. A deliberate reiterated use of a perception module, for instance of the face recognition module when looking for a specific face in a crowd, is also slow and effortful without this implying in any way that the basic mechanism involved is non-modular.

It is tempting at this stage to equate system 1 reasoning with intuitive inferences and system 2 reasoning with reflective inferences. Some analogies seem obvious. Both system 2 and reflective inference are characterised by control, effortfulness, explicitness and, (at least virtual) domain-generality. They contrast in all these respects with system 1 and with intuitive inference.

There are also important disanalogies between the two ways of partitioning inferences. To begin with, the intuitive/reflective contrast is not one between two systems operating at the same level. Intuitive inferences are the direct output of many different modules. Reflective inferences are an indirect output of one of these modules. Hence, there are two ways to pell out the contrast. One may contrast the whole cognitive system, which delivers intuitive inferences through its many component subsystems, with one of these subsystems—the argumentation module—which, like all other inferential modules, directly delivers intuitive inferences, but which also indirectly delivers reflective inferences. One may also contrast a variety of processes carried out by different modules in many different ways, with the processes carried out in a more systematic way by a single one of these modules. These two ways of spelling out the contrast are of course compatible. They highlight the clear asymmetry between a first type of inferences—system 1 or intuitive—found in all animals endowed with rich enough cognitive systems, and a second type of inferences—system 2 or reflective—that may well be absent in non-human animals and that, even in humans, are used much more sparingly than the first type.

The argumentation module, being an ordinary module (different from other modules just as every module is different from all the others), shares to a greater or lesser extent many properties with many other modules. In particular, rather than being unique in requiring high attention (at least when its operations are reiterated in an inferential chain), the argumentation module may just stand towards one end of a particular gradient on which all modules are

situated, a gradient defined by the respective role of bottom-up and top-down triggering factors in the activation of the module.

Some modules—we have mentioned danger detectors—have an inbuilt expectation of relevance and get activated in a bottom-up way. Our rich cognitive lives are possible, however, only to the extent that stimuli capable of pre-empting attention the way danger-detectors do do not occur too frequently in our environment. The full activation of most modules depends on a combination of bottom-up and top-down factors of attention. Consider for instance the detection of mood on the basis of facial cues. In ordinary social life, we are often surrounded by many people with different moods, but typically we pay attention only to the moods of some individuals who particularly matter to us, either in a relatively permanent way, or because of our interaction with them at that moment. So, modern humans can live in an urban environment and encounter new faces continually without having their attention preempted by these stimuli, even if they have dedicated modules to interpret them.

The fact that the full activation of modules depends to a greater or lesser extent on top-down control of attention does not mean at all that we choose consciously which modules to activate. After all, we are not even aware of the existence of our mental modules (or else the modularity thesis would not be controversial). What it means is that, by attending in a more or less voluntary way to some possible inputs and not to others, we modify their relative ease of processing (those that are already being attended to requiring less processing effort), thereby increasing their expected relevance (which is an inverse function of expected effort), and thereby the probability of their being fully processed by specific modules. So, the argumentation module is not unique in being much less dependent on bottom-up than on top-down factors of activation, even if it is likely to be towards one end of this particular gradient. One consequence of this relative and indirect controllability, is that the argumentation module should exhibit greater individual and situational variations than modules at the other extreme of the gradient. Another consequence of controllability is that the outputs of the argumentation module are particularly likely to be consciously attended.

As we already suggested, outputs of modules—unlike modular processes—may be conscious. They are particularly likely to be so when the process of which they are an output results from controlled attention, as in the case of reasoning proper. Moreover, since reasoning takes the form of a series of inferential steps the output of each of which consists in a justification for inferring a given conclusion, reasoning may appear to consciousness as a series of epistemically justified operations performed on explicit representations. This conscious representation of reasoning is partly misleading: what appears to consciousness is

at best the series of intermediate and final conclusions with their justification, that is, a derivation in the formal, abstract sense of the term of the ultimate conclusion from the initial premises, and not a derivation in the concrete sense of a process with a series of sub-processes, (a distinction underscored by (Harman, 1986). Moreover, quite often, some of the intermediate steps may not be consciously entertained at all, so that the derivation is in fact enthymematic. Our approach thus clarifies in what interesting but limited sense reflective inference, or 'system 2 reasoning', may seem conscious, in contrast with intuitive inference or 'system 1 reasoning'. This approach does not suggest that 'consciousness' as such enables reflective inference or plays a causal role in it.

The function of reflective inference

General considerations and experimental evidence give us good reasons to distinguish two types of inferential processes, be they described as system 1 and system 2 reasoning or as intuitive and reflective inference. Still, one would want such a distinction to provide more novel theoretical insights, allow more new analyses of available evidence, and suggest more ground-breaking experimental research than it has done so far.

In this search for greater theoretical and empirical import, it should be fruitful to consider the different functions of these two types of inference, and in particular of reflective inference (the general function of intuitive inference is, we take it, better understood; interesting issues at that level have to do rather with the function of individual domain-specific modules). We assume that these forms of inference are evolved capacities, or at least capacities for the development of which there is an evolved disposition. Ernst Mayr's oft quoted remark is relevant here (extending his point from physiology to psychology): "The adaptationist question, 'What is the function of a given structure or organ?' has been for centuries the basis for every advance in physiology." (Mayr, 1983, p.328). So what, if anything, made the structures underlying different forms of inference advantageous over evolutionary time? The founders of dual system theory have proposed at best cursory answers to this question.

In his 1996 article, Steven Sloman alluded to the question of the two systems' function. He offered two suggestions: the first is that "the systems serve complementary functions. The associative system is able to draw on statistical structure, whereas a system that specializes in analysis and abstraction is able to focus on relevant features." The second suggestion draws

on Freudian psychology: on the one hand, the pain principle motivates us to seek gratification and avoid pain (system 1); on the other hand, we sometimes have to repress these impulses because gratification would otherwise escape us (system 2).

In their book on dual system theory, Jonathan Evans and David Over state that "consciousness [i.e. system 2 reasoning] gives us the possibility to deal with novelty and to anticipate the future" (Evans & Over, 1996, p.154). In a more recent article, Evans says that "interesting though such [evolutionary] speculations are, they may seem to have little immediate relevance to thinking and reasoning researchers attempting to account for the results of their experiments" (Evans, 2006).

In *The Robot's Rebellion*, Keith Stanovich offers a somewhat more elaborate evolutionary account. He states that system 2 is "where the genes gave up direct control and instead said (metaphorically, by the types of phenotypic effect that they created) 'things will be changing too fast out there, brain, for us to tell you exactly what to do—you just go ahead and do what you think is best given the general goals (survival, sexual reproduction) that we have inserted.'" (Stanovich, 2004). He claims that system 1 reasoning is built by our genes to serve them directly by way of contextualized rules of the form: when in situation X, do Y (because it tends to maximize fitness). By contrast, system 2 reasoning is built to favour the individual. To that end, it should be able to fight some of the urges of system 1, by being decontextualized (so that it can find solutions that are not so context-dependent) and by having a strong inhibitory power.

These three views on the evolution and function of system 2 reasoning concur in seeing it as a way to compensate for some of the shortcomings of system 1 and to enhance individual cognition. This is consistent with the view of classical philosophers, Descartes in particular, according to which reasoning (by which they mean conscious reasoning) is the only reliable way to acquire knowledge.

There are, however, strong reasons to doubt that conscious or system 2 reasoning—here we will call it 'reasoning' tout court—evolved to enhance individual cognition. The view that its function is to permit delaying gratification is puzzling for two reasons. The ability to delay gratification when it is advantageous to do so is a widespread feature of animal cognition—in hoarding food for instance—and not a specifically human trait. In humans, the ability to delay gratification seems to be a personality trait related to emotions and is dissociated from the ability to reason, as illustrated by the famous story of Phineas Gage and other better-documented similar cases discussed by Damasio (Damasio, 1994).

The view that the function of reasoning is to enhance the ability to deal with novelty is not compelling either. Humans tend to accumulate in memory information of no immediate practical relevance which they can exploit to imagine possible novel situations. This is a more plausible basis for the ability to deal with novelty. The role of reasoning proper, as opposed to intuitive inference, in memory and imagination can hardly be described as central.[4]

The 'Cartesian' view that reasoning is *the* road to knowledge, or the more cautious view that the function of reasoning is to enhance cognition are also questionable. The issue is one of costs and benefits: those of reasoning have to be compared with those intuitive inference. All theorists agree that reasoning is a relatively slow and costly process. Moreover, reasoning is difficult and prone to a variety of performance errors. So how might such a fallible and costly system still be advantageous? By providing a check on the inferences of system 1? System 1 inferences are on the whole reliable, and it remains to be demonstrated that checking them by means of reasoning, i.e. correcting some mistakes at a high cost and at the risk of further mistakes, would be advantageous. Is reasoning advantageous by allowing the mind to go where it would not intuitively? Many such extensions of the domain of knowledge that make a crucial use of reasoning come to mind, in the sciences in particular, but they typically involve social procedures and institutions where only few individuals make ground-breaking contributions. It is unclear that, at the individual level, the value of reasoning lies in its opening new intellectual vistas.

As an alternative to the view that the basic function of reasoning is to enhance individual cognition, we want to explore the hypothesis that reasoning has a primarily social function and, more specifically, that it is linked to the massive human reliance on communicated information.

Communication is found in a large number of species. For communication to evolve, the cost for the communicator of emitting a signal, and the cost for the receiver of responding to the signal must, on average, be inferior to the benefits. Often, however, the interests of the communicator and of the receiver do not coincide: communicators commonly have an interest in deceiving, whereas receivers' interest is best served by reliable, honest signals. If dishonest signals were frequent to the point of making communication disadvantageous to receivers, receivers would stop responding to them, and emitting these signals would cease to be

[4] There is one kind of contingency that does call for a special form of reasoning, and that is strategic planning in social interaction. According to the Machiavellian hypothesis (Byrne & Whiten, 1988; Whiten & Byrne, 1997), this is in fact a driving force in the evolution of mindreading. To what extent it has evolved as a distinct module (or submodule of mindreading) is an open question. While strategic thinking has features in common with

advantageous to the communicators too (Krebs & Dawkins, 1984). In other words, communication would be selected out. There is a variety of mechanisms that ensure honest signalling in animal species. The signal may involve a cost that only an honest signaller is in a position to incur (Zahavi & Zahavi, 1997). A peacock, for instance, signals its fitness to peahens by displaying a magnificent tail, the cost of which could not be supported by an unfit peacock. Individuals may store information about past communication events and cease to trust communicators that have proved unreliable. Several species of monkey, for instance, are known to recognize the vocalizations of different members of their group; if an individual 'cries wolf' (by using a vocalization in a context that does not warrant it), the other members will soon stop reacting to this individual's vocalizations, or at least to the specific vocalization that has been improperly used (Cheney & Seyfarth, 1990; Gouzoules, Gouzoules, & Miller, 1996).

Non-human animals, however, communicate only very simple information about a narrow range of matters, so that simple ad hoc mechanisms may evolve to enforce honesty. Humans, on the other hand, communicate complex information on an unbounded variety of matters and rely much more on communicated information than any other species. This reliance, hugely advantageous as it may be, is also a source of vulnerability to misinformation and deception. In other terms, there has been among humans a strong selective pressure for ways to filter communicated information so as to come as near as possible to accepting all and only reliable information. We assume that this pressure has caused not one but a variety of mechanisms of what may be called 'epistemic vigilance' (Sperber, Clément, Mascaro, Mercier, Origgi & Wilson, submitted) to evolve. Some of the mechanisms have to do with selectively trusting or distrusting different sources of information on the basis of what is otherwise known of their competence and benevolence towards their audience, their past record in communication, and even behavioural indices of honesty or dishonesty (even if these are only marginally reliable; see (DePaulo et al., 2003; Ekman, 2001). Other mechanisms have to do with properties of the information communicated which make it more or less credible.

A possible way to help calibrate one's trust is for the receiver to check the coherence of what is being said against his own knowledge base. The communicator would then have to adjust her signals if she wants them to be accepted. One way is to stay within the boundaries

---

standard reasoning, in particular its metarepresentational complexity, we suggest that it is the work of a module other than the argumentation module.

of what the receiver will be willing to accept on trust.[5] In some cases however, the communicator might want to communicate some information that the receiver will not accept on trust. Here a strategy for the communicator can be to show the receiver how the information she communicates is in fact coherent with what he already believes—how, in fact, it would be incoherent for him not to accept it. To show this, the communicator must present information that the receiver is already disposed to accept or is willing to accept on trust and that provides premises from which the less easily accepted information follows.

The communicator can moreover highlight the logical or evidential links between the acceptable premises and the intended conclusion. The receiver will then be in a position to evaluate the strength or the validity of these links. Why would he put in such effort? First, it should be noted that most of the effort will be on the communicator's side: it is advantageous for her to make her argument as plain, simple and understandable as possible if she wants to convince the receiver by this means. But the receiver also has something to gain from using a more selective, finer-grained filtering mechanism. Communicators whose benevolence or competence in the matter at hand cannot be taken for granted may nevertheless have valuable information to transmit; it is useful in such cases to be able to bypass or overcome selective distrust. Also, because one's previously held beliefs may be wrong, it can be useful to be able to go beyond a simple check of coherence with these beliefs.

If this scenario is correct, there may have been selection pressures favouring the evolution of capacities that allow communicators and receivers to evaluate evidential, logical and coherence relationships between different pieces of information, i.e. selection pressures for reflective inference. Reflective inference so understood is geared to deal with specific problems concerning the acceptance or rejection of claims in communication. The effectiveness of the argumentation module, like that of any other module, should depend on the expected relevance of its operations in a given situation. In particular, situations characterised by the need to convince others or by that of not being too easily convinced should trigger more efficient reasoning—a prediction quite specific to this approach.

[5] To put it another way: when a trusted communicator communicates something that is not totally coherent with the receiver's beliefs, the receiver has to choose between revising his beliefs regarding the content of what is being communicated or revising his trust in the speaker. He will tend to choose the solution that brings less incoherence, and this will often be to lower trust in the speaker.

Reasoning and argumentation: some evidence

Abstract vs. argumentative contexts

If our approach is right, reasoning should be more easily triggered in argumentative situations. We should therefore expect better performances on reasoning tasks where the participants are placed in such situations. Standard theories make no such prediction. If anything, argumentative contexts should increase the cognitive load since they involve taking into account different opinions.

It is now well established that performances on computationally trivial logical problems can be dismal. As Evans states in his review of the literature on deductive reasoning: “it must be said that logical performance in abstract reasoning tasks is generally quite poor” (Evans, 2002, p. 981). The simplest way to compare the abstract context of a classical reasoning experiment with an argumentative context is to get the participants to discuss the problem in groups.

Among the great many studies on group decision making, the most relevant are those bearing on problems that have a demonstrably correct answer—and are thus analogous to the tasks used in most reasoning experiments. It has now been repeatedly shown that, provided certain minimal conditions are met (the good answer must be accessible to at least one of the participants for instance), what is observed is that in such contexts, if one of the participants has the correct answer, then the other members will get to it too. This has been shown for mathematical tasks (Laughlin & Ellis, 1986; Stasson, Kameda, Parks, Zimmerman, & Davis, 1991), ‘Eureka’ problems (in which the correct solution seems obvious in retrospect—Laughlin, Kerr, Davis, Halff, & Marciniak, 1975), and Mastermind problems (from the board game—Bonner, Baumann, & Dalal, 2002). In all these cases the performance of groups tends to be at the level of the best participants taken individually. The experiments carried out by Moshman and Geil (1998) illustrate this point dramatically. The experimenters had participants solve the Wason selection task, either first individually and then in groups, or directly in groups. In both cases, the performance of the groups was impressively higher than that of the participants who were solving the problem individually: 75% of the groups found the right answer, compared to 14% in the solitary condition[6].

According to the theory advocated here, this dramatic improvement is due to the fact that when they have to solve the problem in groups, participants have to argue and debate, and

that this activates their reasoning abilities in such a manner that they are able either to come up with the correct solution, or at least to accept it and reject the incorrect ones.

Of course, this is not the only possible interpretation of these results. An alternative interpretation might be that the smartest participant gets it right and the others recognize her competence and accept her answer without reasoning (explanation hinted at by (Oaksford, Chater, & Grainger, 1999). Another possible interpretation is that the participants are simply sharing information, and not reasoning. These explanations are hard to reconcile with the following facts. First of all, information sharing is often insufficient to solve the task. For example, in the Wason selection task, it will often be the case that a participant has wrongly selected a card and another has rightly rejected it. In that case, sharing information won't do the trick: participants have conflicting pieces of information, and they have to pick the correct one. This means that conflicts and debate should occur. An analysis of the transcripts of such experiments will show that such is indeed the case (Moshman & Geil, 1998; Trognon, 1993), and there is a large literature showing that conflict is often the crucial factor that allows groups to outperform individuals (see the references in Schulz-Hardt, Brodbeck, Mojzisch, Kerschreiter, & Frey, 2006). In some cases, conflicts will even lead a group in which no individual had the correct answer towards it—provided that not everyone makes the same mistake to start with (this happened in some of the groups studied by Moshman and Geil and is known in developmental psychology as "two wrongs make a right" (Glachan & Light, 1982; Schwarz, Neuman, & Biezuner, 2000) and as the "assembly bonus effect" in social psychology (Kerr, Maccoun, & Kramer, 1996). The explanation based on the recognition of an expert is also hard to reconcile with the presence and importance of such conflicts. One could even argue that the opposite in fact happens: a person is recognised as an expert because she uses good arguments—so participants must use reasoning to discern good arguments in the first place (see Littlepage & Mueller, 1997 for evidence in that direction).

Finally, we can also rule out an explanation based on general motivation: one might think that participants are more motivated—will make greater effort—to solve any task in group. This would be quite surprising, however, given the importance of social loafing in groups (Karau & Williams, 1993). Moreover, if motivation was the problem, it should be alleviated by monetary incentives. However, in line with the general observation that money tends to have no effect on performances in decision making tasks (Camerer & Hogarth, 1999), it has been shown that monetary incentives do not increase the performance in the Wason

[6] A similar effect—if a bit less dramatic—was observed by (Maciejovsky & Budescu, 2007).

selection task (Johnson-Laird & Byrne, 2002; Jones & Sugden, 2001)—a result in sharp contrast with the dramatic improvement in group settings.

Biases in reasoning

The biases that plague reasoning provide further evidence in favour of our approach. We concentrate here on two twin biases that have been reported time and again: the confirmation bias and the disconfirmation bias. Both biases apply when we have to evaluate a belief or a hypothesis: instead of objectively evaluating it, we seek to confirm it if we agree with it in the first place, and to disconfirm it if we don't. This can hardly be sanctioned by a normative theory and is all the more disquieting in that it seems to be extremely widespread: 'smart' people do it (Stanovich & West, 2007), open-minded people do it (Stanovich & West, 2007), physicians, judges, scientists do it (see Fugelsang & Dunbar, 2005; Nickerson, 1998 and references within). From an argumentative viewpoint, these biases are hardly surprising. In fact, they could be predicted on the grounds that when we try to persuade someone that something is true (or false), a confirmation (or disconfirmation) bias may help us achieve our goal.

The experiment that has done the most to promote the idea of the confirmation bias is Wason's 2,4,6 task (Wason, 1960). In this task, participants have to find the rule governing the formation of triplets of numbers, knowing that 2,4,6 is such a triplet. They can test their hypothesis by proposing their own triplets and being told whether they fit the rule. Participants can propose triplets and suggest tentative rules until they have found the correct rule or have given up. Participants typically show a strong confirmation bias in using triplets that conform to their hypothesis in order to test for it rather than triplets that might falsify it (Tweney, Doherty, Warner, & Pliske, 1980; Wason, 1960). Whether the strategy of the participant is really non-normative has been debated (Klayman & Ha, 1987; Koehler, 1993), but two elements point to a real bias and fit with our hypothesis. First, "psychological experiments that have strongly instructed participants to take a falsification approach to the 2,4,6 task have in fact had little effect in improving performance (Poletiek, 1996; Tweney, Doherty, Warner, & Pliske, 1980)" (Evans, 2006). If the tendency to confirm wasn't a deep-rooted bias, these instructions should be much more effective. Even more interestingly, there seems to exist a simple solution to get participants to use a falsifying strategy: when told that

they were testing someone else's hypothesis, participants used the falsifying strategy four times more often and abandoned the hypothesis sooner (Cowley & Byrne, 2005).[7]

Just as we tend to confirm claims we agree with, we tend to disconfirm claims that don't fit our views. The classical demonstration of this disconfirmation bias comes from a study by Lord, Ross and Lepper (1979) in which people had to evaluate studies either in favour or against the death penalty. Participants who supported the death penalty were much more critical of the study arguing against it, and conversely. It has later been shown that people not only put in more effort in examining studies whose conclusions they don't agree with: they are strongly biased towards critical thoughts (Edwards & Smith, 1996). Here, too, urging participants to be objective isn't very efficient (Lord, Lepper, & Preston, 1984). However, when participants are told to imagine that a given study they agree with has in fact the opposite conclusion, they become quite adept at critically examining it (Lord, Lepper, & Preston, 1984).

Another well-known effect in the psychology of reasoning is the belief bias, thought to be a consequence of the disconfirmation bias. Some experiments have pitted the believability of the conclusion of an argument against its logical validity. The main effect is one of believability: people will tend to use believability instead of validity to judge the argument, a finding easy to interpret in terms of epistemic vigilance: believability provides a reason to believe and unbelievability a reason to disbelieve. This effect is stronger for believable conclusions: with them, validity barely makes a difference. Validity, on the other hand, is taken into greater account in th case of unbelievable conclusions. It seems then that when the conclusion of an argument is believable, participants hardly bother to evaluate its validity, but when it is unbelievable, they try to find the flaw, as they should if what motivates is epistemic vigilance. When they fail to find any flaw because the argument is valid, they accept it, again, as they should. Epistemic vigilance is what explains that people are more sensitive to validity when the conclusion is unbelievable (Evans, Newstead, & Byrne, 1993; Klauer, Musch, & Naumer, 2000; Newstead, Pollard, Evans, & Allen, 1992).

In a nutshell, the confirmation, disconfirmation, and believability biases behave much as the argumentative theory would predict: one cannot suppress them with instructions to be objective, but if one can get participants to change their mind about the claim to be evaluated, then the biases can disappear or even be reversed.

[7] Even though due to the limited number of participants this difference failed to reach significance.

Are people good at arguing?

If argumentation has been so important in evolutionary history, then humans should be good at it. The first large-scale study of the "skills of argument" done by Deanna Kuhn (1991) concluded, however, that people are rather poor at argumentation. We remain unconvinced for the following reasons. To begin with, the context of her experiments was quite artificial: people were asked to argue about topics of which they had very limited knowledge (e.g. the causes of school failure or of relapse into delinquency after prison) with an experimenter who wasn't really arguing with them. It has later been shown that when participants are more knowledgeable about the topics they stop making some of the mistakes noted by Kuhn (such as using explanation—some kind of naive theory— instead of genuine evidence (Brem & Rips, 2000). More importantly, most of the shortcomings Kuhn attributed to participants were in fact instances of the confirmation bias: for instance, participants often had trouble finding alternative theories or rebuttals to their own hypothesis. So they were indeed doing what one should expect if reasoning was used not to get at the truth but to persuade.

Other studies have since tended to show that participants do possess the skills necessary to understand and take part in an argument. They can follow the commitments of the different speakers and determine, at any given point of the argument, who has the burden of proof (Bailenson & Rips, 1996; Rips, 1998). They understand the macrostructure of arguments (Ricco, 2003). They are often able to spot the classical fallacies of argumentation, such as *ad hominem*, *petitio principii* (begging the question), or circular reasoning (Neuman, Glassner, & Weinstock, 2004; Neuman, Weinstock, & Glasner, 2006; Rips, 2002; Weinstock, Neuman, & Tabak, 2004).

Note that these 'fallacies' can sometimes be quite appropriate – for example, when someone uses her authority to make a point, then a good *ad hominem* may be effective. Oaksford, Hahn and their colleagues have used Bayesian statistics to pinpoint which features of a given argument make it more or less fallacious (Hahn & Oaksford, 2007). For example, the validity of a slippery slope argument depends—among other things—on the conditional probability of each step of the slope given the preceding step. Instead of the mere ability to spot fallacies, a more accurate measure of people's argumentative skills is a measure of the fit between their evaluations of arguments and their actual validity (as indexed by these statistics). In a set of experiments, these researchers tested these predictions for a set of such 'fallacies': argument from ignorance (Oaksford & Hahn, 2004), slippery slope arguments and circular reasoning (Hahn & Oaksford, 2006). In all of these experiments, participants' ratings

of the strength of different arguments were indeed correlated with factors that reflected the actual statistical validity of the arguments.

Finally, researchers who have looked at real arguments—between participants debating for instance—have been "impressed by the coherence of the reasoning displayed. Participants … appear to build complex argument and attack structure. People appear to be capable of recognizing these structures and of effectively attacking their individual components as well as the argument as a whole" (Resnick, Salmon, Zeitz, Wathen, & Holowchak, 1993). The contrast between these observations and the dismal results of simple reasoning tasks could not be sharper.

Conclusion

In this chapter we have tried to outline an original view of reasoning, seeing it as an aspect of social, and more specifically communicative competence. This view is embedded in an evolutionary psychology framework and in particular in a massive-modularist view of the human mind. At its core is the distinction between two types of inferences: intuitive inferences that are the direct output of inferential modules and take place without attention to reasons for accepting them; and reflective inferences that are an indirect output of a particular metarepresentational module, the argumentation module, the direct output of which is an argument for or against a given conclusion. Being a distinction between two types of inferences, this view has obvious analogies with other dual process accounts of reasoning, but some serious differences were also noted.

An evolutionary argument was put forward to explain what the function of the argumentation module might be—namely, to regulate the flow of information among interlocutors through persuasiveness on the side of the communicator and epistemic vigilance on the side of the audience. Testable predictions follow from such an account. We have argued that some puzzling findings in the psychology of reasoning and state-of-the-art work in the psychology of argumentation confirm these predictions. Further confirmation will have to come from novel experiments specifically designed to test these theoretical claims.

Acknowledgments: We thank Jonathan Evans, Keith Frankish, Katherine Kinzler and Deirdre Wilson for their useful comments. This work was made possible in part thanks to the support of the *to be completed*

References

Bailenson, J. N., & Rips, L. J. (1996). Informal reasoning and burden of proof. *Applied Cognitive Psychology, 10*(7), 3-16.

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, Massachusetts: MIT Press.

Barrett, H. C. (2005). Enzymatic computation and cognitive modularity. *Mind and language, 20*(3), 259-287.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review, 113*(3), 628-647.

Bonnay, D., & Simmenauer, B. (2005). Tonk strikes back. *Australasian Journal of Logic, 3*, 33–44.

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision making and performance. *Organizational Behavior and Human Decision Processes, 88*, 719–736.

Braine, M. D. S. (1990). The "natural logic" approach to reasoning. In W. F. Overton (Ed.), *Reasoning, Necessity and Logic: Developmental Perspectives* (Vol. 133-157). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24*, 573–604.

Byrne, R., & Whiten, A. (Eds.). (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford: Oxford University Press.

Camerer, C., & Hogarth, R. M. (1999). The effect of financial incentives on performance in experiments: a review and capital-labor theory. *Journal of Risk and Uncertainty, 19*, 7-42.

Carruthers, P. (2006) *The Architecture of the Mind*. Oxford: Oxford University Press.

Cheney, D. L., & Seyfarth, R. M. (1990). *How Monkeys See the World*. Chicago: Chicago University Press.

Cohen, L. J. (1992). *An Essay on Belief and Acceptance*. Oxford: Clarendon Press.

Cowley, M., & Byrne, R. M. J. (2005). *When falsification is the only path to truth.* Paper presented at the Twenty-Seventh Annual Conference of the Cognitive Science Society, Stresa, Italy.

Damasio, A. R. (1994). *Descartes' Error: Emotion Reason, and the Human Brain*. New York: GP Putnam's Sons.

de Sousa, R. (1971). How to give a piece of your mind: or, the logic of belief and assent. *Review of Metaphysics, 25*(1), 52-79.

Dennett, D. C. (1969). *Content and Consciousness*. London: Routledge and Kegan Paul.

Dennett, D. C. (1981). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, Mass.: MIT Press.
DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychol Bull, 129*(1), 74-118.
Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71*, 5-24.
Ekman, P. (2001). *Telling Lies*. New York: Norton.
Engel, P. (2000). *Believing and Accepting*. Dordrecht: Kluwer Academic Publishers.
Engel, P. (2006). Logic, reasoning and the logical constants. *Croatian Journal of Philosophy*(2 (17)), 219.
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. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454-459.
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. (in press). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*.
Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.
Evans, J. S. B. T., & Over, D. E. (1996). *Rationality and Reasoning*. Hove: Psychology Press.
Fodor, J. (2001). *The Mind Doesn't Work That Way*. Cambridge, Massachusetts: MIT Press.
Frankish, K. (this volume) Systems and levels: Dual-system theories and the personal-subpersonal distinction
Fugelsang, J. A., & Dunbar, K. N. (2005). Brain-based mechanisms underlying complex causal thinking. *Neuropsychologia, 43*(8), 1204-1213.
Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.
Glachan, M., & Light, P. (1982). Peer interaction and learning: Can two wrongs make a right? In G. Butterworth & P. Light (Eds.), *Social cognition: Studies in the development of understanding* (pp. 238–262). Chicago: University of Chicago Press.
Gouzoules, H., Gouzoules, S., & Miller, K. (1996). Skeptical responding in rhesus monkeys (Macaca mulatta). *International Journal of Primatology, 17*, 549-568.
Hahn, U., & Oaksford, M. (2006). A bayesian approach to informal argument fallacies. *Synthese, 152*(2), 207-236.
Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological Review, 114*(3), 704-73229.
Harman, G. (1986). *Change in View: Principles of Reasoning*. Cambridge, Mass.: MIT Press.
Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge, UK: Cambridge University Press.
Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109*, 646-678.
Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision, 50*(1), 59-99.
Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–294). Cambridge, UK: Cambridge Univiversity Press.
Karau, S. J., & Williams, K. D. (1993). Social loafing: a meta-analytic review and theoretical integration. *Journal of personality and social psychology, 65*(4), 681-706.
Kerr, N. L., Maccoun, R. J., & Kramer, G. P. (1996). Bias in judgement: comparing individuals and groups. *Psychological review, 103*(4), 687-719.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychol Rev, 107*(4), 852-884.
Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.
Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes, 56*, 28-55.
Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (2ème ed., pp. 390-402). Oxford: Basil Blackwell Scientific Publications.
Kuhn, D. (1991). *The Skills of Arguments*. Cambridge: Cambridge 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.
Laughlin, P. R., Kerr, N. L., Davis, J. H., Halff, H. M., & Marciniak, K. A. (1975). Group size, member ability, and social decision schemes on an intellective task. *Journal of Personality and Social Psychology, 33*, 80-88.
Leslie, A. M. (1995). A theory of agency. In D. Sperber & D. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate*. New York: Oxford University Press.
Littlepage, G. E., & Mueller, A. L. (1997). Recognition and utilization of expertise in problem-solving groups: Expert characteristics and behavior. *Group Dynamics, 1*, 324-328.
Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*, 1231-1243.
Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098-2109.
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.
Mayr, E. (1983). How to carry out the adaptationist program. *The American Naturalist, 121*(3), 324-334.
Moshman, D., & Geil, M. (1998). Collaborative Reasoning: Evidence for Collective Rationality. *Thinking and Reasoning, 4*(3), 231-248.
Neuman, Y., Glassner, A., & Weinstock, M. (2004). The effect of a reason's truth-value on the judgment of a fallacious argument. *Acta Psychologica, 116*(2), 173-184.
Neuman, Y., Weinstock, M. P., & Glasner, A. (2006). The effect of contextual factors on the judgement of informal reasoning fallacies. *The Quarterly Journal of Experimental Psychology, 59*(2), 411-425.
Newstead, S. E., Pollard, P., Evans, J. S. B. T., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition, 45*, 257-284.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology, 2*, 175-220.
Oaksford, M., Chater, N., & Grainger, R. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5*, 193-243.
Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology, 58*(2), 75-85.
Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin and Review, 11*(6), 988-1010.

Poletiek, F. H. (1996). Paradoxes of falsification. *Quarterly Journal of Experimental Psychology, 49A*, 447-462.
Prior, A. N. (1960). The runabout inference-ticket. *Analysis, 21*(2), 38-39.
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.
Ricco, R. B. (2003). The macrostructure of informal arguments: A proposed model and analysis. *The Quarterly Journal of Experimental Psychology A, 56*(6), 1021-1051.
Rips, L. J. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA: MIT Press.
Rips, L. J. (1998). Reasoning and conversation. *Psychological Review, 105*, 411–441.
Rips, L. J. (2002). Circular reasoning. *Cognitive Science, 26*, 767–795.
Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., & Frey, D. (2006). Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *Journal of personality and social psychology, 91*(6), 1080-1093.
Schwarz, B. B., Neuman, Y., & Biezuner, S. (2000). Two wrongs make a right. . .if they argue together! *Cognition and Instruction, 18*, 461–494.
Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059-1074.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3-22.
Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*(1), 29-56.
Sperber, D. (1985). *On Anthropological Knowledge*: Cambridge University Press Cambridge.
Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 39-67). Cambridge: Cambridge University Press.
Sperber, D. (1997). Intuitive and reflective beliefs. *Mind and Language, 12*(1), 67-83.
Sperber, D. (2000a). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.), *Metarepresentations: A Multidisciplinary Perspective* (pp. 117-137). Oxford: Oxford University Press.
Sperber, D. (2001a). An evolutionary perspective on testimony and argumentation. *Philosophical Topics, 29*, 401-413.
Sperber, D. (2001b). In defense of massive modularity. In E. Dupoux (Ed.), *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler* (pp. 47-57). Cambridge, Massachusetts: MIT Press.
Sperber, D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence & S. Stich (Eds.), *The Innate Mind: Structure and Contents*.
Sperber, D. (Ed.). (2000b). *Metarepresentations: A Multidisciplinary Perspective*. Oxford: Oxford University Press.
Sperber, D., Clément, F., Mascaro, O., Mercier, H., Origgi, G. & Wilson, D. (submitted). Epistemic vigilance.
Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.
Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind and Language, 17*, 3-23.
Stalnaker, R. C. (1984). *Inquiry*. Cambridge, Mass: MIT Press
Stanovich, K. E. (2004). *The Robot's Rebellion*. Chicago: Chicago University Press.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645-726.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning, 13*(3), 225-247.

Stasson, M. F., Kameda, T., Parks, C. D., Zimmerman, S. K., & Davis, J. H. (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly, 54*, 25-35.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, Massachusetts: Harvard University Press.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides & J. Tooby (Eds.), *The Adapted Mind* (pp. 19-136). Oxford: Oxford University Press.

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.

Tweney, R. D., Doherty, M. E., Warner, W. J., & Pliske, D. B. (1980). Strategies of rule discovery in an inference task. *Quaterly Journal of Experimenal Psychology, 32*, 109-124.

Vickers, J. M. (1988). *Chance and Structure: An Essay on the Logical Foundations of Probability*. Oxford: Clarendon (Oxford University Press).

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-137.

Weinstock, M., Neuman, Y., & Tabak, I. (2004). Missing the point or missing the norms? Epistemological norms as predictors of students' ability to identify fallacious arguments. *Contemporary Educational Psychology, 29*(1), 77-94.

Whiten, A., & Byrne, R. W. (Eds.). (1997). *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge: Cambridge University Press.

Wilson, D. (2000). Metarepresentation in linguistic communication. In D. Sperber (Ed.), *Metarepresentations: A Multidisciplinary Perspective*. Oxford: Oxford University Press.

Wynn, K. (1992). Addition and Subtraction in Human Infants. *Nature, 358*, 749-750.

Zahavi, A., & Zahavi, A. (1997). *The handicap principle: a missing piece of Darwin's puzzle*. Oxford: Oxford University Press.