Why do humans reason? Arguments for an argumentative theory

Hugo Mercier
Philosophy, Politics and Economics Program, University of Pennsylvania,
Philadelphia, PA 19104
hmercier@sas.upenn.edu
http://hugo.mercier.googlepages.com/

Dan Sperber
Jean Nicod Institute (EHESS-ENS-CNRS), 75005 Paris, France
and Department of Philosophy, Central European University, Budapest, Hungary
dan@sperber.fr
http://www.dan.sperber.fr

Abstract: Reasoning is generally seen as a means to improve knowledge and make better decisions. However, much evidence shows that reasoning often leads to epistemic distortions and poor decisions. This suggests that the function of reasoning should be rethought. Our hypothesis is that the function of reasoning is argumentative. It is to devise and evaluate arguments intended to persuade. Reasoning so conceived is adaptive given the exceptional dependence of humans on communication and their vulnerability to misinformation. A wide range of evidence in the psychology of reasoning and decision making can be reinterpreted and better explained in the light of this hypothesis. Poor performance in standard reasoning tasks is explained by the lack of argumentative context. When the same problems are placed in a proper argumentative setting, people turn out to be skilled arguers. Skilled arguers, however, are not after the truth but after arguments supporting their views. This explains the notorious confirmation bias. This bias is apparent not only when people are actually arguing but also when they are reasoning proactively from the perspective of having to defend their opinions. Reasoning so motivated can distort evaluations and attitudes and allow erroneous beliefs to persist. Proactively used reasoning also favors decisions that are easy to justify but not necessarily better. In all these instances traditionally described as failures or flaws, reasoning does exactly what can be expected of an argumentative device: Look for arguments that support a given conclusion, and, ceteris paribus, favor conclusions for which arguments can be found.

Keywords: argumentation; confirmation bias; decision making; dual process theory; evolutionary psychology; motivated reasoning; reason-based choice; reasoning
Inference (as the term is most commonly understood in psychology) is the production of new mental representations on the basis of previously held representations.

Examples of inferences are the production of new beliefs on the basis of previous beliefs, the production of expectations on the basis of perception, or the production of plans on the basis of preferences and beliefs. So understood, inference need not be deliberate or conscious. It is at work not only in conceptual thinking but also in perception and in motor control (Kersten et al. 2004; Wolpert & Kawato 1998). It is a basic ingredient of any cognitive system. Reasoning, as commonly understood, refers to a very special form of inference at the conceptual level, where not only is a new mental representation (or conclusion) consciously produced, but the previously held representations (or premises) that warrant it are also consciously entertained. The premises are seen as providing reasons to accept the conclusion. Most work in the psychology of reasoning is about reasoning so understood. Such reasoning is typically human. There is no evidence that it occurs in nonhuman animals or in preverbal children.¹

How do humans reason? Why do they reason? These two questions are mutually relevant, since the mechanisms for reasoning should be adjusted to its function. While the how-question has been systematically investigated (e.g., Evans et al. 1993; Johnson-Laird 2006; Oaksford & Chater 2007; Rips 1994), there is very little discussion of the why-question. How come? It may be that the function of reasoning is considered too obvious to deserve much attention. According to a long philosophical tradition, reasoning is what enables the human mind to go beyond mere perception, habit, and instinct. In the first, theoretical section of this article we sketch
a tentative answer to the how-question and then focus on the why-question: We outline an approach to reasoning based on the idea that the primary function for which it evolved is the production and evaluation of arguments in communication. In sections 2–5, we consider some of the main themes and findings in the experimental literature on reasoning and show how our approach helps make better sense of much of the experimental evidence and hence gains empirical support from it.

1. Reasoning: Mechanism and function

1.1. Intuitive inference and argument

Since the 1960s, much work in the psychology of reasoning has suggested that, in fact, humans reason rather poorly, failing at simple logical tasks (Evans 2002), committing egregious mistakes in probabilistic reasoning (Kahneman & Tversky 1972; Tversky & Kahneman 1983), and being subject to sundry irrational biases in decision making (Kahneman et al. 1982). This work has led to a rethinking of the mechanisms for reasoning, but not – or at least not to the same degree – of its assumed function of enhancing human cognition and decision making. The most important development has been the emergence of dual-process models that distinguish between intuitions and reasoning (or system 1 and system 2 reasoning) (Evans 2007; Johnson-Laird 2006; Kahneman 2003; Kahneman & Frederick 2002; 2005; Sloman 1996; Stanovich 2004). Here we outline our own dual-process approach: We contend in particular that the arguments used in reasoning are the output of a mechanism of intuitive inference (Mercier & Sperber 2009; Sperber 1997; 2001).
A process of inference is a process the representational output of which necessarily or probabilistically follows from its representational input. The function of an inferential process is to augment and correct the information available to cognitive system. An evolutionary approach suggests that inferential processes, rather than being based on a single inferential mechanism or constituting a single integrated system, are much more likely to be performed by a variety of domain-specific mechanisms, each attuned to the specific demands and affordances of its domain (e.g., see Barkow et al. 1992). The inferential processes carried out by these mechanisms are unconscious: They are not mental acts that individuals decide to perform, but processes that take place inside their brain, at a “sub-personal” level (in the sense of Dennett 1969). People may be aware of having reached a certain conclusion – be aware, that is, of the output of an inferential process – but we claim that they are never aware of the process itself. All inferences carried out by inferential mechanisms are in this sense intuitive. They generate intuitive beliefs; that is, beliefs held without awareness of reasons to hold them.

The claim that all inferential processes carried out by specialized inferential mechanisms are unconscious and result in intuitive inferences may seem to contradict the common experience of forming a belief because one has reflected on reasons to accept it, and not, or not only, because of its intuitive force. Such beliefs, held with awareness of one’s reasons to hold them, are better described not as intuitive but as reflective beliefs (Sperber 1997). Our consciously held reason for accepting a reflective belief may be trust in its source (the professor, the doctor, the priest). Our reasons may also have to do with the content of the belief: We realize, for instance,
that it would be inconsistent on our part to hold to our previous beliefs and not accept some given new claim. Far from denying that we may arrive at a belief through reflecting on our reasons to accept it, we see this as *reasoning proper*, the main topic of this article. What characterizes reasoning proper is indeed the awareness not just of a conclusion but of an argument that justifies accepting that conclusion. We suggest, however, that arguments exploited in reasoning are the output of an intuitive inferential mechanism. Like all other inferential mechanisms, its processes are unconscious (as also argued by Johnson-Laird 2006, p. 53; and Jackendoff 1996) and its conclusions are intuitive. However, *these intuitive conclusions are about arguments*; that is, about representations of relationships between premises and conclusions.

The intuitive inferences made by humans are not only about ordinary objects and events in the world. They can also be about representations of such objects or events (or even about higher-order representations of representations). The capacity to represent representations, and to draw inferences about them, is a *metarepresentational* capacity with formal properties relevant to the mental computations involved (Recanati 2000; Sperber 2000b). Several mental mechanisms use this metarepresentational capacity. In particular, humans have a mechanism for representing mental representations and for drawing intuitive inferences about them. This Theory of Mind mechanism is essential to our understanding of others and of ourselves (Leslie 1987; Premack & Woodruff 1978). Humans also have a mechanism for representing verbal representations and for drawing intuitive inferences about
them. This pragmatic mechanism is essential to our understanding of communicated meaning in context (Grice 1975; Sperber & Wilson 2002).

We want to argue that there is yet another intuitive metarepresentational mechanism, a mechanism for representing possible reasons to accept a conclusion – that is, for representing arguments – and for evaluating their strength. Arguments should be sharply distinguished from inferences. An inference is a process the output of which is a representation. An argument is a complex representation. Both an inference and an argument have what can be called a conclusion, but in the case of an inference, the conclusion is the output of the inference; in the case of an argument, the conclusion is a part – typically the last part – of the representation. The output of an inference can be called a “conclusion” because what characterizes an inferential process is that its output is justified by its input; the way however in which the input justifies the output is not represented in the output of an intuitive inference. What makes the conclusion of an argument a “conclusion” (rather than simply a proposition) is that the reasons for drawing this conclusion on the basis of the premises are (at least partially) spelled out. As Gilbert Harman (1986) has justly argued, it is a common but costly mistake to confuse the causally and temporally related steps of an inference with the logically related steps of an argument. The causal steps of an inference need not recapitulate the logical step of any argument for it to be an inference, and the logical step of an argument need not be followed in any inference for it to be an argument.

Descartes’s famous Cogito argument, “I think therefore I am,” illustrates the manner in which an argument can be the output of an intuitive inference. Most people
believe intuitively that they exist, and are not looking for reason to justify this belief. But should you look for such reasons – that is, should you take a reflective stance towards the proposition that you exist – Descartes’s argument would probably convince you: It is intuitively evident that the fact that you are thinking is a good enough reason to accept that you exist, or, in other terms, that it would be inconsistent to assert “I think” and to deny “I am.” What is not at all obvious in this particular case are the reasons for accepting that this intuitively good argument is truly a good argument, and philosophers have been hotly debating the issue (e.g., Katz 1986).

Simple as the Cogito or more complex, all arguments must ultimately be grounded in intuitive judgments that given conclusions follow from given premises. In other words, we are suggesting that arguments are not the output of a system 2 mechanism for explicit reasoning, that would be standing apart from, and in symmetrical contrast to, a system 1 mechanism for intuitive inference. Rather, arguments are the output of one mechanism of intuitive inference among many that delivers intuitions about premise-conclusion relationships. Intuitions about arguments have an evaluative component: Some arguments are seen as strong, others as weak. Moreover there may be competing arguments for opposite conclusions and we may intuitively prefer one to another. These evaluation and preferences are ultimately grounded in intuition.

If we accept a conclusion because of an argument in its favor that is intuitively strong enough, this acceptance is an epistemic decision that we take at a personal level. If we construct a complex argument by linking argumentative steps, each of which we see as having sufficient intuitive strength, this is a personal-level mental
action. If we verbally produce the argument so that others will see its intuitive force and will accept its conclusion, it is a public action that we consciously undertake. The mental action of working out a convincing argument, the public action of verbally producing this argument so that others will be convinced by it, and the mental action of evaluating and accepting the conclusion of an argument produced by others correspond to what is commonly and traditionally meant by *reasoning* (a term that can refer to either a mental or a verbal activity).

Why should the reflective exploitation of one mechanism for intuitive inference among many stand out as so important that it has been seen as what distinguishes humans from beasts? Why, in dual-process theories of reasoning, should it be contrasted on its own with all the mechanisms for intuitive inference taken together? We see three complementary explanations for the saliency of reasoning. First, when we reason, we know that we are reasoning, whereas the very existence of intuitive inference was seen as controversial in philosophy before its discovery in cognitive science. Second, while an inferential mechanism that delivers intuitions about arguments is, strictly speaking, highly domain specific, the arguments that it delivers intuitions about can be representations of anything at all. Thus, when we reason on the basis of these intuitions, we may come to conclusions in all theoretical and practical domains. In other words, even though inferences about arguments are domain specific (as evolutionary psychologists would expect), they have domain general consequences and provide a kind of virtual domain generality (without which traditional and dual-process approaches to reasoning would make little sense). Third,
as we will now argue, the very function of reasoning puts it on display in human communication.

1.2. The function of reasoning

We use function here in its biological sense (see Allen et al. 1998). Put simply, a function of a trait is an effect of that trait that causally explains its having evolved and persisted in a population: Thanks to this effect, the trait has been contributing to the fitness of organisms endowed with it. In principle, several effects of a trait may contribute to fitness, and hence a trait may have more than a single function. Even then, it may be possible to rank the importance of different functions, and in particular to identify a function for which the trait is best adapted as its main function. For instance, human feet have the functions of allowing us both to run and to walk, but their plantigrade posture is better adapted for walking than for running, and this is strong evidence that walking is their main function (Cunningham et al. 2010). In the same vein, we are not arguing against the view that our reasoning ability may have various advantageous effects, each of which may have contributed to its selection as an important capacity of the human mind. We do argue, however, that reasoning is best adapted for its role in argumentation, which should therefore be seen as its main function.

There have been a few tentative attempts in dual-process approaches to explain the function and evolution of reasoning. The majority view seems to be that the main function of reasoning is to enhance individual cognition. This is expressed, for instance, by Kahneman (2003, p. 699), Gilbert (2002), Evans and Over (1996, p.
Stanovich (2004, p. 64), and Sloman (1996, p. 18). This classical view of reasoning – it goes back to Descartes and to ancient Greek philosophers – faces several problems that become apparent when its functional claims are laid out in slightly greater detail. It is sometimes claimed (e.g., by Kahneman 2003) that the meliorative function of system 2 reasoning is achieved by correcting mistakes in system 1 intuitions. However, reasoning itself is a potential source of new mistakes. Moreover, there is considerable evidence that, when reasoning is applied to the conclusions of intuitive inference, it tends to rationalize them rather than to correct them (e.g., Evans & Wason 1976).

According to another hypothesis, conscious reasoning “gives us the possibility to deal with novelty and to anticipate the future” (Evans & Over 1996, p. 154). But giving an organism the possibility to deal with novelty and to anticipate the future is less a characterization of reasoning than it is of learning (or even, it could be argued, of cognition in general). After all, learning can be defined as “the process by which we become able to use past and current events to predict what the future holds” (Niv & Schoenbaum 2008, p. 265). The issue is not whether, on occasion, reasoning can help correct intuitive mistakes or better adapt us to novel circumstances. No doubt, it can. The issue is how far these occasional benefits explain the costs incurred, and hence the very existence of reasoning among humans, and also explain its characteristic features. In any case, evolutionary hypotheses are of little help unless precise enough to yield testable predictions and explanations. To establish that reasoning has a given function, we should be able at least to identify signature effects of that function in the very way reasoning works.
Here we want to explore the idea that the emergence of reasoning is best understood within the framework of the evolution of human communication. Reasoning enables people to exchange arguments that, on the whole, make communication more reliable and hence more advantageous. The main function of reasoning, we claim, is argumentative (Sperber 2000a; 2001; see also Billig 1996; Dessalles 2007; Kuhn 1992; Perelman & Olbrechts-Tyteca 1969; Haidt 2001 and Gibbard 1990 offer a very similar take on the special case of moral reasoning.

For communication to be stable, it has to benefit both senders and receivers; otherwise they would stop sending or stop receiving, putting an end to communication itself (Dawkins & Krebs 1978; Krebs & Dawkins 1984). But stability is often threatened by dishonest senders who may gain by manipulating receivers and inflicting too high of a cost on them. Is there a way to ensure that communication is honest? Some signals are reliable indicators of their own honesty. Costly signals such as a deer antlers or a peacock tail both signal and show evidence that the individual is strong enough to pay that cost (Zahavi & Zahavi 1997). Saying “I am not mute” is proof that the speaker is indeed not mute. However, for most of the rich and varied informational contents that humans communicate among themselves, there are no available signals that would be proof of their own honesty. To avoid being victims of misinformation, receivers must therefore exercise some degree of what may be called epistemic vigilance (Sperber et al., 2010). The task of epistemic vigilance is to evaluate communicator and the content of their messages in order to filter communicated information.
Several psychological mechanisms may contribute to epistemic vigilance. The two most important of these mechanisms are trust calibration and coherence checking. People routinely calibrate the trust they grant different speakers on the basis of their competence and benevolence (Petty & Wegener 1998). Rudiments of trust calibration based on competence have been demonstrated in 3-year-old children (for reviews, see Clément, in press; Harris 2007). The ability to distrust malevolent informants has been shown to develop in stages between the ages of 3 and 6 (Mascaro & Sperber 2009).

The interpretation of communicated information involves activating a context of previously held beliefs and trying to integrate the new with old information. This process may bring to the fore incoherencies between old and newly communicated information. Some initial coherence checking thus occurs in the process of comprehension. When it uncovers some incoherence, an epistemically vigilant addressee must choose between two alternatives. The simplest is to reject communicated information, thus avoiding any risk of being misled. This may, however, deprive the addressee of valuable information and of the opportunity to correct or update earlier beliefs. The second, more elaborate, alternative consists in associating coherence checking and trust calibration and allowing for a finer-grained process of belief revision. In particular, if a highly trusted individual tells us something that is incoherent with our previous beliefs, some revision is unavoidable: We must revise either our confidence of the source or our previous beliefs. We are likely to choose the revision that reestablishes coherence at the lesser cost, and this will often consist in accepting the information communicated and revising our beliefs.
What are the options of a communicator wanting to communicate a piece of information that the addressee is unlikely to accept on trust? One option may be for the communicator to provide evidence of her reliability in the matter at hand (for instance, if the information is about health issues, she might inform the addressee that she is a doctor). But what if the communicator is not in a position to boost her own authority? Another option is to try to convince her addressee by offering premises the addressee already believes or is willing to accept on trust, and showing that, once these premises are accepted, it would be less coherent to reject the conclusion than to accept it. This option consists in producing arguments for one’s claims and in encouraging the addressee to examine, evaluate, and accept these arguments.

Producing and evaluating arguments is, of course, a use of reasoning.

Reasoning contributes to the effectiveness and reliability of communication by allowing communicators to argue for their claim and by allowing addressees to assess these arguments. It thus increases both in quantity and in epistemic quality the information humans are able to share. Claiming as we do that this role of reasoning in social interaction is its main function fits well with much current work stressing the role of sociality in the unique cognitive capacities of humans (Byrne & Whiten 1988; R. I. M. Dunbar 1996; R. I. M. Dunbar & Shultz 2003; Hrdy 2009; Humphrey 1976; Tomasello et al. 2005; Whiten & Byrne 1997). In particular, the evolutionary role of small group cooperation has recently been emphasized (Dubreuil, 2010; Sterelny, in press). Communication plays an obvious role in human cooperation both in the setting of common goals and in the allocation of duties and rights. Argumentation is uniquely effective in overcoming disagreements that are likely to occur, in particular...
in relatively equalitarian groups. While there can hardly be any archaeological
evidence for the claim that argumentation already played an important role in early
human groups, we note that anthropologists have repeatedly observed people arguing
in small-scale traditional societies (Boehm et al. 1996; Brown 1991; Mercier, in
press).

The main function of reasoning is argumentative: Reasoning has evolved and
persisted mainly because it makes human communication more effective and
advantageous. As most evolutionary hypotheses, this claim runs the risk of being
perceived as another “just so story.” It is therefore crucial to show that it entails
falsifiable predictions. If the main function of reasoning is indeed argumentative, then
it should exhibit as signature effects strengths and weaknesses related to the relative
importance of this function compared to other potential functions of reasoning. This
should be testable through experimental work done here and now. Our goal now is to
spell out and explain what signature effects we predict, to evaluate these predictions
in light of the available evidence, and to see whether they help make better sense of a
number of well-known puzzles in the psychology of reasoning and decision making.
Should one fail, on the other hand, to find such signature of the hypothesized
argumentative function of reasoning, and even more should one find that the main
features of reasoning match some other function, then our hypothesis should be
considered falsified.²

Several predictions can be derived from the argumentative theory of
reasoning. The first and most straightforward is that reasoning should do well what it
evolved to do; that is, produce and evaluate arguments (sects. 2.1 and 2.2). In general,
adaptations work best when they are used to perform the task they evolved to perform. Accordingly, reasoning should produce its best results when used in argumentative contexts, most notably in group discussions (sect. 2.3). When we want to convince an interlocutor with a different viewpoint, we should be looking for arguments in favor of our viewpoint rather than in favor of hers. Therefore, the next prediction is that reasoning used to produce arguments should exhibit a strong confirmation bias (sect. 3). A further related prediction is that, when people reason on their own about one of their opinions, they are likely to do so proactively, that is, anticipating a dialogic context, and mostly to find arguments that support their opinion. Evidence of the existence of such motivated reasoning is reviewed in section 4. Finally, we want to explore the possibility that, even in decision making, the main function of reasoning is to produce arguments to convince others rather than to find the best decision. Thus, we predict that reasoning will drive people towards decisions for which they can argue – decisions that they can justify – even if these decisions are not optimal (sect. 5).

2. Argumentative skills

2.1. Understanding and evaluating arguments

In this section, we review evidence showing that people are skilled arguers, using reasoning both to evaluate and to produce arguments in argumentative contexts. This, in itself, is compatible with other accounts of the main function of reasoning. However, this evidence is relevant because the idea that people are not very skilled arguers is relatively common; if it were true, then the argumentative theory would be
a nonstarter. It is therefore crucial to demonstrate that this is not the case and that people have good argumentative skills, starting with the ability to understand and evaluate arguments.

The understanding of arguments has been studied in two main fields of psychology: persuasion and attitude change, on the one hand, and reasoning, on the other. The aims, methods, and results are different in the two fields. Within social psychology, the study of persuasion and attitude change has looked at the effects of arguments on attitudes. In a typical experiment, participants hear or read an argument (a “persuasive message”), and the evolution of their attitude on the relevant topic is measured. For instance, in a classic study by Petty and Cacioppo (1979), participants were presented with arguments supporting the introduction of a comprehensive senior exam. Some participants heard strong arguments (such as data showing that “graduate and professional schools show a preference for undergraduates who have passed a comprehensive exam”), while others heard much weaker arguments (such as a quote from a graduate student saying that “since they have to take comprehensives, undergraduates should take them also”). In this experiment, it was shown that participants who would be directly affected by the setting up of a comprehensive exam were much more influenced by strong arguments than by weak ones. This experiment illustrates the more general finding stemming from this literature that, when they are motivated, participants are able to use reasoning to evaluate arguments accurately (for a review, see Petty & Wegener 1998).

The demonstration that people are skilled at assessing arguments seems to stand in sharp contrast with findings from the psychology of reasoning. In a typical
reasoning experiment, participants are presented with premises and asked either to produce or to evaluate a conclusion that should follow logically. Thus, they may have to determine what, if anything, follows from premises such as “If there is a vowel on the card, then there is an even number on the card. There is not an even number on the card.” In such tasks, Evans (2002) recognizes that “logical performance . . . is generally quite poor” (p. 981). To give just one example, it was found in a review that an average of 40% of participants fail to draw the simple *modus tollens* conclusion that was used as an example (if *p* then *q*, not *q*, therefore not *p*) (Evans et al. 1993).

However, reasoning, according to the present view, should mostly provide a felicitous evaluation in dialogic contexts – when someone is genuinely trying to convince us of something. This is not the case in these decontextualized tasks that involve no interaction or in abstract problems. In fact, as soon as these logical problems can be made sense of in an argumentative context, performance improves. For instance, participants can easily understand a *modus tollens* argument when it is of use not simply to pass some test but to evaluate communicated information (see Thompson et al. 2005b); the production of valid *modus tollens* arguments in argumentative contexts is also “surprisingly common” (Pennington & Hastie 1993, p. 155).

While students of reasoning focus on logical fallacies, other scholars have turned to the study of the fallacies of argumentation. Unlike logical fallacies, fallacies of argumentation come in degrees: Depending on their content and context, they can be more or less fallacious. For instance, a *slippery-slope fallacy* (where a claim is criticized for being a step on a slope that ends up with a blatant mistake) is in fact
valid to the extent that, having made the first step on the slope, it is probable that one will continue all the way down (Corner et al. 2006).

Various experiments have shown that participants are generally able to spot other argumentative fallacies (Hahn & Oaksford 2007, experiment 3; Neuman 2003; Neuman et al. 2006; Weinstock et al. 2004; see also Corner & Hahn 2009). Not only do they spot them, but they tend to react appropriately: rejecting them when they are indeed fallacious, or being convinced to the degree that they are well grounded (Corner et al. 2006; Hahn & Oaksford 2007; Hahn et al. 2005; Oaksford & Hahn 2004; Rips 2002). When researchers have studied other skills specific to argumentation, performance has proved to be satisfactory. Thus, participants are able to recognize the macrostructure of arguments (Ricco 2003), to follow the commitments of different speakers (Rips 1998), and to attribute the burden of proof appropriately (Bailenson & Rips 1996; see also Rips 1998, experiment 3). On the whole, the results reviewed in this section demonstrate that people are good at evaluating arguments both at the level of individual inferences and at the level of whole discussions.

2.2. Producing arguments

The first studies that systematically investigated argument production used the following methodology. Participants were asked to think about a given topic, such as “Would restoring the military draft significantly increase America’s ability to influence world events?” (Perkins 1985) or “What are the causes of school failure?” (Kuhn 1991). After being left to think for a few minutes, they had to state and defend
their view to the experimenter. The conclusions of these studies were quite bleak and
highlighted three main flaws. The first is that people resort to mere explanations
(“make sense” causal theories) instead of relying on genuine evidence (data) to
support their views. However, later research has shown that this is mostly an artifact
of the lack of evidence available to the participants: When evidence is made
available, participants will favor it (in both production and evaluation) (Brem & Rips
2000; see also Hagler & Brem 2008; Sá et al. 2005). A second flaw noted by Perkins
and Kuhn is the relative superficiality of the arguments used by participants. This can
be explained by a feature of the tasks: Unlike in a real debate, the experimenter didn’t
challenge the arguments of the participants, however weak they were. In a normal
argumentative setting, a good argument is an argument that is not refuted. As long as
they are not challenged, it makes sense to be satisfied with seemingly superficial
arguments. On the other hand, people should be able to generate better arguments
when engaged in a real debate. This is exactly what Kuhn and her colleagues
observed: Participants who had to debate on a given topic showed a significant
improvement in the quality of the arguments they used afterwards (Kuhn et al. 1997;
for similar results with analogical reasoning, see Blanchette & Dunbar 2001).

The third flaw, according to Perkins and Kuhn, is the most relevant one here.
Participants had generally failed to anticipate counterarguments and generate
rebuttals. For these two authors, and indeed the critical thinking tradition, this is a
very serious failing. Seen from an argumentative perspective, however, this may not
be a simple flaw but rather a feature of argumentation that contributes to its
effectiveness in fulfilling its function. If one’s goal is to convince others, one should
be looking first and foremost for supportive arguments. Looking for
counterarguments against one’s own claims may be part of a more sophisticated and
effortful argumentative strategy geared to anticipating the interlocutor’s response,
but, in the experimental setting, there was no back-and-forth to encourage such an
extra effort (and participants knew not to expect such a back-and-forth). If this is a
correct explanation of what need not be a flaw after all, then the difficulty that people
seem to have in coming up with counterarguments should be easily overcome by
having them challenge someone else’s claims rather than defending their own.
Indeed, when mock jurors were asked to reach a verdict and were then presented with
an alternative verdict, nearly all of them were able to find counterarguments against it
(Kuhn et al. 1994). In another experiment, all participants were able to find
counterarguments against a claim (which was not theirs) and to do so very quickly
(Shaw 1996).

When people have looked at reasoning performance in felicitous
argumentative settings, they have observed good results. Resnick and her colleagues
(1993) created groups of three participants who disagreed on a given issue. Analyzing
the debates, the researchers were “impressed by the coherence of the reasoning
displayed. Participants . . . appear to build complex arguments 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” (pp. 362–
63; see also Blum-Kulka et al. 2002; Hagler & Brem 2008; Stein & Bernas 1997;
Stein et al. 1996). It is worth noting that a strikingly similar pattern emerges from
developmental studies (see Mercier, submitted a).
To sum up, people can be skilled arguers, producing and evaluating arguments felicitously. This good performance stands in sharp contrast with the abysmal results found in other, nonargumentative, settings, a contrast made particularly clear by the comparison between individual and group performance.

2.3. Group reasoning

If people are skilled at both producing and evaluating arguments, and if these skills are displayed most easily in argumentative settings, then debates should be especially conducive to good reasoning performance. Many types of tasks have been studied in group settings, with very mixed results (for recent reviews, see Kerr & Tindale 2004; Kerr et al. 1996). The most relevant findings here are those pertaining to logical or, more generally, intellective tasks “for which there exists a demonstrably correct answer within a verbal or mathematical conceptual system” (Laughlin & Ellis 1986, p. 177). In experiments involving this kind of task, participants in the experimental condition typically begin by solving problems individually (pretest), then solve the same problems in groups of four or five members (test), and then solve them individually again (posttest), to ensure that any improvement does not come simply from following other group members. Their performance is compared to those of a control group of participants who take the same tests but always individually.

Intellective tasks allow for a direct comparison with results from the individual reasoning literature, and the results are unambiguous. The dominant scheme (Davis 1973) is truth wins, meaning that, as soon as one participant has understood the problem, she will be able to convince the whole group that her solution is correct
This can lead to big improvements in performance. Some experiments using the Wason selection task dramatically illustrate this phenomenon (Moshman & Geil 1998; see also Augustinova 2008; Maciejovsky & Budescu 2007). The Wason selection task is the most widely used task in reasoning, and the performance of participants is generally very poor, hovering around 10% of correct answers (Evans 1989; Evans et al. 1993; Johnson-Laird & Wason 1970). However, when participants had to solve the task in groups, they reached the level of 80% of correct answers.

Several challenges can be leveled against this interpretation of the data. It could be suggested that the person who has the correct solution simply points it out to the others, who immediately accept it without argument, perhaps because they have recognized this person as the “smartest” (Oaksford et al. 1999). The transcripts of the experiments show that this is not the case: Most participants are willing to change their mind only once they have been thoroughly convinced that their initial answer was wrong (e.g., see Moshman & Geil 1998; Trognon 1993). More generally, many experiments have shown that debates are essential to any improvement of performance in group settings (for a review and some new data, see Schulz-Hardt et al. 2006; for similar evidence in the development and education literature, see Mercier, submitted a). Moreover, in these contexts, participants decide that someone is smart based on the strength and relevance of her arguments and not the other way around (Littlepage & Mueller 1997). Indeed, it would be very hard to tell who is “smart” in such groups – even if general intelligence were easily perceptible, it only correlates .33 with success in the Wason selection task (Stanovich & West 1998).
Finally, in many cases, no single participant had the correct answer to begin with. Several participants may be partly wrong and partly right, but the group will collectively be able to retain only the correct parts and thus converge on the right answer. This leads to the *assembly bonus effect* in which the performance of the group is better than that of its best member (Blinder & Morgan 2000; Laughlin et al. 2002; 2003; 2006; Lombardelli et al. 2005; Michaelsen et al. 1989; Sniezek & Henry 1989; Stasson et al. 1991; Tindale & Sheffey 2002). Once again there is a striking convergence here, with the developmental literature showing how groups – even when no member had the correct answer initially – can facilitate learning and comprehension of a wide variety of problems (Mercier, submitted a).

According to another counterargument, people are simply more motivated, generally, when they are in groups (Oaksford et al. 1999). This is not so. On the contrary, “the ubiquitous finding across many decades of research (e.g., see Hill 1982; Steiner 1972) is that groups usually fall short of reasonable potential productivity baselines” (Kerr & Tindale 2004, p. 625). Moreover, other types of motivation have no such beneficial effect on reasoning. By and large, monetary incentives, even substantial ones, fail to improve performance in reasoning and decision-making tasks (Ariely et al., 2009; Bonner & Sprinkle 2002; Bonner et al. 2000; Camerer & Hogarth 1999; and, in the specific case of the Wason selection task, see Johnson-Laird & Byrne 2002; Jones & Sugden, 2001). Thus, not any incentive will do: Group settings have a motivational power to which reasoning responds specifically.
The argumentative theory also helps predict what will happen in nonoptimal group settings. If all group members share an opinion, a debate should not arise spontaneously. However, in many experimental and institutional settings (juries, committees), people are forced to discuss, even if they already agree. When all group members agree on a certain view, each of them can find arguments in its favor. These arguments will not be critically examined, let alone refuted, thus providing other group members with additional reasons to hold that view. The result should be a strengthening of the opinions held by the group (for a review, see Sunstein 2002; for a recent illustration, see Hinsz et al. 2008). Contra Sunstein’s law of group polarization, it is important to bear in mind that this result is specific to artificial contexts in which people debate even though they tend to agree in the first place. When group members disagree, discussions often lead to depolarization (Kogan & Wallach 1966; Vinokur & Burnstein 1978). In both cases, the behavior of the group can be predicted on the basis of the direction and strength of the arguments accessible to group members, as demonstrated by research carried out in the framework of the Persuasive Argument Theory (Vinokur 1971), which ties up with the prediction of the present framework (Ebbesen & Bowers 1974; Isenberg 1986; Kaplan & Miller 1977; Madsen 1978).

The research reviewed in this section shows that people are skilled arguers: They can use reasoning both to evaluate and to produce arguments. This good performance offers a striking contrast with the poor results obtained in abstract reasoning tasks. Finally, the improvement in performance observed in argumentative
settings confirms that reasoning is at its best in these contexts. We will now explore in more depth a phenomenon already mentioned in this section: the confirmation bias.

3. The confirmation bias: A flaw of reasoning or a feature of argument production?

The confirmation bias consists in the “seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand” (Nickerson 1998, p. 175). It is one of the most studied biases in psychology (for review, see Nickerson 1998). While there is some individual variation, it seems that everybody is affected to some degree, irrespective of factors like general intelligence or open mindedness (Stanovich & West 2007; 2008a; 2008b). For standard theories of reasoning, the confirmation bias is no more than a flaw of reasoning. For the argumentative theory, however, it is a consequence of the function of reasoning and hence a feature of reasoning when used for the production of arguments.

In fact, we suggest, the label confirmation bias has been applied to two distinct types of case, both characterized by a failure to look for counterevidence or counterarguments to an existing belief, both consistent with the argumentative approach but brought about in different ways. In cases that deserve the label of confirmation bias, people are trying to convince others. They are typically looking for arguments and evidence to confirm their own claim, and ignoring negative arguments and evidence unless they anticipate having to rebut them. While this may be seen as a bias from a normative epistemological perspective it clearly serves the goal of convincing others. In another type of case, we are dealing not with biased reasoning
but with an absence of reasoning proper. Such an absence of reasoning is to be expected when people already hold some belief on the basis of perception, memory, or intuitive inference, and do not have to argue for it. Say, I believe that my keys are in my trousers because this is where I remember putting them. Time has passed, and they could now be in my jacket, for instance. However, unless I have some positive reason to think otherwise, I just assume that they are still in my trousers, and I don’t even make the inference (which, if I am right, would be valid) that they are not in my jacket or any of the other places where, in principle, they might be. In such cases, people typically draw positive rather than negative inferences from their previous beliefs. These positive inferences are generally more relevant to testing these beliefs. For instance, I am more likely to get conclusive evidence that I was right or wrong by looking for my keys in my trousers rather than in my jacket (even if they turn out not to be in my jacket, I might still be wrong in thinking that they are in my trousers). We spontaneously derive positive consequences from our intuitive beliefs. This is just a trusting use of our beliefs, not a confirmation bias (see Klayman & Ha 1987).

The theory we are proposing makes three broad predictions. The first is that the genuine confirmation bias (as opposed to straightforward trust in one’s intuitive beliefs and their positive consequences) should occur only in argumentative situations. The second is that it should occur only in the production of arguments. The rationale for a confirmation bias in the production of arguments to support a given claim does not extend to the evaluation of arguments by an audience that is just aiming to be well informed. The third prediction is that the confirmation bias in the production of arguments is not a bias in favor of confirmation in general and against
disconfirmation in general: It is a bias in favor of confirming one’s own claims, which should be naturally complemented by a bias in favor of disconfirming opposing claims and counterarguments.

3.1. Hypothesis testing: No reasoning, no reasoning bias

One of the areas in which the confirmation bias has been most thoroughly studied is that of hypothesis testing, often using Wason’s rule discovery task (Wason 1960). In this task, participants are told that the experimenter has in mind a rule for generating number triples and that they have to discover it. The experimenter starts by giving participants a triple that conforms to the rule (2, 4, 6). Participants can then think of a hypothesis about the rule and test it by proposing a triple of their own choice. The experimenter says whether or not this triple conforms to the rule. Participants can repeat the procedure until they feel ready to put forward their hypothesis about the rule. The experimenter tells them whether or not their hypothesis is true. If it is not, they can try again or give up.

Participants overwhelmingly propose triples that fit with the hypothesis they have in mind. For instance, if a participant has formed the hypothesis “three even numbers in ascending order,” she might try 8, 10, 12. As argued by Klayman and Ha (1987), such an answer corresponds to a “positive test strategy” of a type that would be quite effective in most cases. This strategy is not adopted in a reflective manner, but is rather, we suggest, the intuitive way to exploit one’s intuitive hypotheses, as when we check that our keys are where we believe we left them as opposed to
checking that they are not where it follows from our belief that they should not be.

What we see here, then, is a sound heuristic rather than a bias.

This heuristic misleads participants in this case only because of some very peculiar (and expressly designed) features of the task. What is really striking is the failure of attempts to get participants to reason in order to correct their ineffective approach. It has been shown that, even when instructed to try to falsify the hypotheses they generate, fewer than one participant in ten is able to do so (Poleti 1996; Tweney et al. 1980). Since the hypotheses are generated by the participants themselves, this is what we should expect in the current framework: The situation is not an argumentative one and does not activate reasoning. However, if a hypothesis is presented as coming from someone else, it seems that more participants will try to falsify it and will give it up much more readily in favor of another hypothesis (Cowley & Byrne 2005). The same applies if the hypothesis is generated by a minority member in a group setting (Butera et al. 1992). Thus, falsification is accessible provided that the situation encourages participants to argue against a hypothesis that is not their own.

3.2. The Wason selection task

A similar interpretation can be used to account for results obtained with the Wason selection task (Wason 1966). In this task, participants are given a rule describing four cards. In the original version, the cards have a number on one side and a letter on the other, although only one side is visible – they might see, for instance, 4, E, 7, and K. The rule might read, “If there is a vowel on one side, then there is an even number on
The task is to say what cards need to be turned over to determine whether the rule is true. In this task, too, it is useful to distinguish the effects of intuitive mechanisms from those of reasoning proper (as has long been suggested by Wason & Evans 1975). Intuitive mechanisms involved in understanding utterances will draw the participants’ attention to the cards that are made most relevant by the rule and the context (Girotto et al. 2001; Sperber et al. 1995). In the standard case, these will simply be the cards mentioned in the rule (the vowel E and the even number 4), as opposed to those that would yield the correct answer (the E and the 7). Given that the 4 can only confirm the rule but not falsify it, the behavior of participants who select this card could be interpreted as showing a confirmation bias. However, as first discovered by Evans (Evans & Lynch 1973), the simple addition of a negation in the rule (“if there is a vowel on one side, then there is not an even number on the other side”) leaves the answers unchanged (the E and 4 are still made relevant), but in this case these cards correspond to the correct, falsifying, response. So these intuitive mechanisms are not intrinsically linked to either confirmation or falsification: They just happen to point to cards that in some cases might confirm the rule and, in other cases, might falsify it.

Confirmation bias does occur in the selection task but at another level. Once the participants’ attention has been drawn to some of the cards, and they have arrived at an intuitive answer to the question, reasoning is used not to evaluate and correct their initial intuition but to find justifications for it (Evans 1996; Lucas & Ball 2005; Roberts & Newton 2002). This is a genuine confirmation bias. As with hypothesis testing, this does not mean that participants are simply unable to understand the task.
or to try to falsify the rule – only that an appropriate argumentative motivation is lacking. That participants can understand the task is shown by the good performance in group settings, as mentioned earlier. Participants should also be able to try to falsify the rule when their first intuition is that the rule is false and they want to prove it wrong. Researchers have used rules such as “all members of group A are Y,” where Y is a negative or positive stereotype (Dawson et al. 2002). Participants who were most motivated to prove the rule wrong – those belonging to group A when Y was negative – were able to produce more than 50% of correct answers, whereas participants from all the other conditions (groups other than A and/or positive stereotype) remained under 20%.

3.3. Categorical syllogisms

Categorical syllogisms are one of the most studied types of reasoning. Here is a typical example: “No C are B; All B are A; therefore some A are not C.” Although they are solvable by very simple programs (e.g., see Geurts 2003), syllogisms can be very hard to figure out – the one just offered by way of illustration, for instance, is solved by less than 10% of participants (Chater & Oaksford 1999). In terms of the mental model theory, what the participants are doing is constructing a model of the premises and deriving a possible conclusion from it (Evans et al. 1999). This constitutes the participants’ initial intuition. To correctly solve the problem, participants should then try to construct counterexamples to this initial conclusion. But this would mean trying to falsify their own conclusion. The present theory predicts that they will not do so spontaneously. And indeed, “any search for
counterexample models is weak . . . participants are basing their conclusions on the first model that occurs to them” (Evans et al. 1999, p. 1505; see also Klauer et al. 2000; Newstead et al. 1999).

Again, we suggest, this should not be interpreted as revealing a lack of ability but only a lack of motivation. When participants want to prove a conclusion wrong, they will find ways to falsify it.

This happens with normal conclusions presented by someone else (Sacco & Bucciarelli 2008) or when participants are faced with so-called unbelievable conclusions such as “All fish are trout.” In this case, they will try to prove that the premises lead to the logical opposite of the conclusion (“Not all fish are trout”) (Klauer et al. 2000). Given that falsification leads to better answers on these tasks, this explains why participants actually perform much better when the conclusion is unbelievable (e.g., see Evans et al. 1983). It is not that they reason more in this case – they spend as much time trying to solve problems with believable conclusions as with unbelievable ones (Thompson et al. 2005c). It is just that the direction reasoning takes is mostly determined by the participants’ initial intuitions. If they have arrived at the conclusion themselves, or if they agree with it, they try to confirm it. If they disagree with it, they try to prove it wrong. In all cases, what they do is try to confirm their initial intuition.

3.4. Rehabilitating the confirmation bias

In all three cases just reviewed – hypothesis testing, the Wason selection task, and syllogistic reasoning – a similar pattern can be observed. Participants have intuitions
that lead them towards certain answers. If reasoning is used at all, it is mostly used to confirm these initial intuitions. This is exactly what one should expect of an argumentative skill, and so these results bolster our claim that the main function of reasoning is argumentative. By contrast, if people were easily able to abstract from this bias, or if they were subject to it only in argumentative settings, then this would constitute evidence against the present theory.

According to a more standard explanation of the confirmation bias, it is an effect of limitations in cognitive resources and in particular in working memory (e.g. Johnson-Laird 2006). But it is hard to reconcile this explanation with the fact that people are very good at falsifying propositions when they are inclined to disagree with them. In those cases, people are not held back by limited resources even though the tasks are not cognitively easier.

However, the idea that the confirmation bias is a normal feature of reasoning that plays a role in the production of arguments may seem surprising in light of the poor outcomes it has been claimed to cause. Conservatism in science is one example (see Nickerson 1998 and references therein). Another is the related phenomenon of groupthink, which has been held responsible for many disasters, from the Bay of Pigs fiasco (Janis 1982) to the tragedy of the Challenger shuttle (Esser & Lindoerfer 1989; Moorhead et al. 1991) (for review, see Esser 1998). In such cases, reasoning tends not to be used in its normal context: that is, the resolution of a disagreement through discussion. When one is alone or with people who hold similar views, one’s arguments will not be critically evaluated. This is when the confirmation bias is most likely to lead to poor outcomes. However, when reasoning is used in a more felicitous
context – that is, in arguments among people who disagree but have a common interest in the truth – the confirmation bias contributes to an efficient form of division of cognitive labor.

When a group has to solve a problem, it is much more efficient if each individual looks mostly for arguments supporting a given solution. They can then present these arguments to the group, to be tested by the other members. This method will work as long as people can be swayed by good arguments, and the results reviewed in section 2 show that this is generally the case. This joint dialogic approach is much more efficient than one where each individual on his or her own has to examine all possible solutions carefully.5 The advantages of the confirmation bias are even more obvious given that each participant in a discussion is often in a better position to look for arguments in favor of his or her favored solution (situations of asymmetrical information). So group discussions provide a much more efficient way of holding the confirmation bias in check. By contrast, the teaching of critical thinking skills, which is supposed to help us overcome the bias on a purely individual basis, does not seem to yield very good results (Ritchart & Perkins 2005; Willingham 2008).

For the confirmation bias to play an optimal role in discussions and group performance, it should be active only in the production of arguments and not in their evaluation. Of course, in the back-and-forth of a discussion, the production of one’s own arguments and the evaluation of those of the interlocutor may interfere with each other, making it hard to properly assess the two processes independently. Still, the evidence reviewed in section 2.1 on the understanding of arguments strongly suggests
that people tend to be more objective in evaluation than in production. If this were not the case, the success of group reasoning reviewed in section 2.3 would be very hard to explain.

4. Proactive reasoning in belief formation

According to the argumentative theory, reasoning is most naturally used in the context of an exchange of arguments during a discussion. But people can also be proactive and anticipate situations in which they might have to argue to convince others that their claims are true or that their actions are justified. We would say that much reasoning anticipates the need to argue. In this section, we will show that work on motivated reasoning can be usefully reinterpreted in this perspective, and, in the next section, we will show that the same applies to work on reason-based choice.

Many of our beliefs are likely to remain unchallenged because they are relevant only to ourselves and we don’t share them or because they are uncontroversial among the people we interact with or because we have sufficient authority to be trusted when we assert them. While we think of most of our beliefs – to the extent that we think about them at all – not as beliefs but just as pieces of knowledge, we are also aware that some of them are unlikely to be universally shared, or to be accepted on trust just because we express them. When we pay attention to the contentious nature of these beliefs, we typically think of them as opinions. Opinions are likely to be challenged and may have to be defended. It makes sense to look for arguments for our opinions before we find ourselves called upon to state them. If the search for arguments is successful, we will be ready. If not, then perhaps it might be
better to adopt a weaker position, one that is easier to defend. Such uses of reasoning have been intensively studied under the name of *motivated reasoning* (Kunda 1990; see also Kruglanski & Freund 1983; Pyszczynski & Greenberg 1987; for a recent review, see Molden & Higgins 2005).

### 4.1. Motivated reasoning

A series of experiments by Ditto and his colleagues, involving reasoning in the context of a fake medical result, illustrate the notion of motivated reasoning (Ditto & Lopez 1992; Ditto et al. 1998; 2003). Participants had to put some saliva on a strip of paper and were told that, if the strip changed color or did not change color, depending on the condition, this would be an indication of an unhealthy enzyme deficiency. Participants, being motivated to believe they were healthy, tried to garner arguments for this belief. In one version of the experiment, participants were told the rate of false positives, which varied across conditions. The use they made of this information reflects motivated reasoning. When the rate of false positives was high, participants who were motivated to reject the conclusion used it to undermine the validity of the test. This same high rate of false positives was discounted by participants who were motivated to accept the conclusion. In another version of the experiment, participants were asked to mention events in their medical history that could have affected the results of the test, which gave them an opportunity to discount these results. Participants motivated to reject the conclusion listed more such events, and the number of events listed was negatively correlated with the evaluation of the test. In these experiments, the very fact that the participant’s health is being tested indicates
that it cannot be taken for granted. The reliability of the test itself is being discussed. This experiment, and many others to be reviewed in this article, demonstrate also that motivated reasoning is not mere wishful thinking (a form of thinking that, if it were common, would in any case be quite deleterious to fitness and would not be coherent with the present theory). If desires did directly affect beliefs in this way, then participants would simply ignore or dismiss the test. Instead, what they do is look for evidence and arguments to show that they are healthy or at least for reasons to question the value of the test.

Other studies have demonstrated the use of motivated reasoning to support various beliefs that others might challenge. Participants dig in and occasionally alter their memories to preserve a positive view of themselves (Dunning et al. 1989; Ross et al. 1981; Sanitioso et al. 1990). They modify their causal theories to defend some favored belief (Kunda 1987). When they are told the outcome of a game on which they had made a bet, they use events in the game to explain why they should have won when they lost (Gilovich 1983). Political experts use similar strategies to explain away their failed predictions and bolster their theories (Tetlock 1998). Reviewers fall prey to motivated reasoning and look for flaws in a paper in order to justify its rejection when they don’t agree with its conclusions (Koehler 1993; Mahoney 1977). In economic settings, people use information flexibly so as to be able to justify their preferred conclusions or arrive at the decision they favor (Boiney et al. 1997; Hsee 1995; 1996a; Schweitzer & Hsee 2002).

All these experiments demonstrate that people sometimes look for reasons to justify an opinion they are eager to uphold. From an argumentative perspective, they
do this not to convince themselves of the truth of their opinion but to be ready to meet the challenges of others. If they find themselves unprepared to meet such challenges, they may become reluctant to express an opinion they are unable to defend and less favorable to the opinion itself, but this is an indirect individual effect of an effort that is aimed at others. In a classical framework, where reasoning is seen as geared to achieving epistemic benefits, the fact that it may be used to justify an opinion already held is hard to explain, especially since, as we will now show, motivated reasoning can have dire epistemic consequences.

4.2. Consequences of motivated reasoning

4.1.1. Biased evaluation and attitude polarization. In a landmark experiment, Lord and colleagues (1979) asked participants who had been previously selected as being either defenders or opponents of the death penalty to evaluate studies relating to its efficiency as a deterrent. The studies given to the participants had different conclusions: While one seemed to show that the death penalty had a significant deterrent effect, the other yielded the opposite result. Even though the methodologies of the two studies were almost identical, the studies that yielded a conclusion not in line with the participants’ opinions were consistently rated as having been much more poorly conducted. In this case, participants used reasoning not so much to assess the studies objectively as to confirm their initial views by finding either flaws or strengths in similar studies, depending on their conclusion. This phenomenon is known as biased assimilation or biased evaluation. This second description is somewhat misleading. In this experiment – and the many related experiments that
have followed it – participants are indeed asked to evaluate an argument. However, what they do is mostly produce arguments to support or rebut the argument they are evaluating, depending on whether they agree with its conclusion or not. Participants are not trying to form an opinion: They already have one. Their goal is argumentative rather than epistemic, and ends up being pursued at the expense of epistemic soundness. That participants engage in this biased search for arguments even when their task is to evaluate an argument has been demonstrated by the experiments we now describe.

Several other experiments have studied the way people evaluate arguments depending on whether they agree or disagree with the conclusions. When people disagree with the conclusion of an argument, they often spend more time evaluating it (Edwards & Smith 1996). This asymmetry arises from the trivial fact that rejecting what we are told generally requires some justification, whereas accepting it does not. Moreover, the time spent on these arguments is mostly devoted to finding counterarguments (Edwards & Smith 1996; see also Brock 1967; Cacioppo & Petty 1979; Eagly et al. 2000). Participants tend to comb through arguments for flaws and end up finding some, whether they are problems with the design of a scientific study (Klaczynski & Gordon 1996b; Klaczynski & Narasimham 1998; Klaczynski & Robinson 2000), issues with a piece of statistical reasoning (Klaczynski & Gordon 1996a; Klaczynski & Lavallee 2005; Klaczynski et al. 1997), or argumentative fallacies (Klaczynski 1997). In all these cases, motivated reasoning leads to a biased assessment: Arguments with unfavored conclusions are rated as less sound and less persuasive than arguments with favored conclusions.
Sometimes the evaluation of an argument is biased to the point where it has an opposite effect to the one intended by the arguer: On reading an argument with a counterattitudinal conclusion (one that goes against their own beliefs or preferences), interlocutors may find so many flaws and counterarguments that their initial unfavorable attitude is in fact strengthened. This is the phenomenon of attitude polarization, which has been studied extensively since it was first demonstrated by (Lord et al. 1979; see also Greenwald 1969; Pomerantz et al. 1995). Taber and Lodge (2006) have demonstrated that, in the domain of politics, attitude polarization is most easily observed in participants who are most knowledgeable (see also Braman 2009; Redlawsk 2002). Their knowledge makes it possible for these participants to find more counterarguments, leading to more biased evaluations.

4.1.2. Polarization, bolstering, and overconfidence. Attitude polarization can also occur in simpler circumstances. Merely thinking about an object may be enough to strengthen attitudes towards it (polarization). This phenomenon has been repeatedly demonstrated. Sadler and Tesser (1973) had participants listen to a recording of a very pleasant-sounding or unpleasant-sounding individual. They then had to give their opinion of this individual, either after having to think about him or her or after performing a distraction task. As expected, the opinions were more extreme (in both directions) when participants had to think about the individual. Tesser and Conlee (1975) showed that polarization increases with the time spent thinking about an item, and Jellison and Mills (1969) showed that it increases with the motivation to think. As in the case of polarization following biased evaluation, such polarization occurs
only when participants are knowledgeable (Tesser & Leone 1977; see also Millar & Tesser 1986). And the effect can be mitigated by providing a reality check: The simple presence of the target object will dramatically decrease polarization (Tesser 1976).

Some later experiments used a slightly different methodology (Chaiken & Yates 1985; Liberman & Chaiken 1991). Instead of simply thinking about the target object, participants had to write a small essay about it. Not only was polarization observed in this case, but it was correlated with the direction and number of the arguments put forward in the essay. These results demonstrate that reasoning contributes to attitude polarization and strongly suggest that it may be its main factor. When people are asked to think about a given item towards which they intuitively have a positive or negative attitude, what happens, we suggest, is that they reflect less on the item itself than on how to defend their initial attitude. Many other experiments have shown that, once people have formed an attitude to a target, they will look for information that supports this attitude (a phenomenon known as selective exposure; see Hart et al. 2009; Smith et al. 2008) and try to put any information they are given to the same use (Bond et al. 2007; Brownstein 2003), which leads them to choose inferior alternatives (Russo et al. 2006).

According to the argumentative theory, reasoning should be even more biased once the reasoner has already stated her opinion, thereby increasing the pressure on her to justify it rather than moving away from it. This phenomenon is called bolstering (McGuire 1964). Thus, when participants are committed to an opinion, thinking about it will lead to a much stronger polarization (Lambert et al. 1996;
Millar & Tesser 1986). *Accountability* (the need to justify one’s decisions) will also increase bolstering (Tetlock et al. 1989; for review, see Lerner & Tetlock 1999).

Finally, motivated reasoning should also affect confidence. When participants think of an answer to a given question, they will be spontaneously tempted to generate reasons supporting that answer. This may then cause them to be overconfident in the answer. Koriat and his colleagues (1980) have tested this hypothesis by using general knowledge questions such as “the Sabines were part of (a) ancient India or (b) ancient Rome.” After answering the question, participants had to produce reasons relevant to their answers. Some participants were asked to generate reasons supporting their answer, while others were asked for reasons against it. The results for people who were explicitly asked to generate reasons supporting their answer were no different from those in a control condition where no reasons were asked for. This suggests that thinking of reasons to support their answer is what people do spontaneously anyhow when they regard their answer not as an obvious piece of knowledge but as an opinion that might be challenged. By contrast, participants in the other group were much less overconfident. Having to think of arguments against their answer enabled them to see its limitations – something they would not do on their own (for replications and extensions to the phenomenon of hindsight bias and the fundamental attribution error, see Arkes et al. 1988; Davies 1992; Griffin & Dunning 1990; Hirt & Markman 1995; Hoch 1985; Yates et al. 1992). It is then easy to see that overconfidence would also be reduced by having participants discuss their answers with people who favor different conclusions.
4.1.3 Belief perseverance. Motivated reasoning can also be used to hang on to beliefs even when they have been proved to be ill-founded. This phenomenon, known as belief perseverance, is “one of social psychology’s most reliable phenomena” (Guenther & Alicke 2008, p. 706; for an early demonstration, see Ross et al. 1975). The involvement of motivated reasoning in this effect can be demonstrated by providing participants with evidence both for and against a favored belief. If belief perseverance were a simple result of some degree of psychological inertia, then the first evidence presented should be the most influential, whether it supports or disconfirms the favored belief. On the other hand, if evidence can be used selectively, then only evidence supporting the favored belief should be retained, regardless of the order of presentation. Guenther and Alicke (2008) tested this hypothesis in the following way. Participants first had to perform a simple perceptual task. This task, however, was described as testing for “mental acuity,” a made-up construct that was supposed to be related to general intelligence, making the results of the test highly relevant to participant’s self-esteem. Participants were then given positive or negative feedback, but a few minutes later they were told that the feedback was actually bogus and the real aim of the experiment was explained. At three different points, the participants also had to evaluate their performance: right after the task, after the feedback, and after the debriefing. In line with previous results, the participants who had received positive feedback showed a classic belief-perseverance effect and discounted the debriefing, which allowed them to preserve a positive view of their performance. By contrast, those who had received negative feedback did the opposite: They took the debriefing fully into account, which allowed them to reject the negative
feedback and restore a positive view of themselves. This strongly suggests that belief perseverance of the type just described is an instance of motivated reasoning (for applications to the domain of political beliefs, see Prasad et al. 2009).\(^{11}\)

### 4.1.4 Violation of moral norms

The results reviewed so far have shown that motivated reasoning can lead to poor epistemic outcomes. We will now see that our ability to “find or make a reason for everything one has a mind to do” (Franklin 1799) can also allow us to violate our moral intuitions and behave unfairly. In a recent experiment, Valdesolo and DeSteno (2008) have demonstrated the role reasoning can play in maintaining moral hypocrisy (when we judge someone else’s action by using tougher moral criteria than we use to judge our own actions). Here is the basic setup. On arriving at the laboratory, participants were told that they would be performing one of two tasks: a short and fun task or a long and hard task. Moreover, they were given the possibility of choosing which task they would be performing, knowing that the other task would be assigned to another participant. They also had the option of letting a computer choose at random how the tasks would be distributed. Once they were done assigning the tasks, participants had to rate how fair they had been. Other participants, instead of having to make the assignment themselves, were at the receiving end of the allocation and had no choice whatsoever; they had to rate the fairness of the participant who had done the allocation, knowing the exact conditions under which this had been done. It is then possible to compare the fairness ratings of participants who have assigned themselves the easy task with the ratings of those who have been assigned the hard task. The difference between these two ratings is a mark
of moral hypocrisy. The authors then hypothesized that reasoning, since it allows participants to find excuses for their behavior, was responsible for this hypocrisy. They tested this hypothesis by replicating the above conditions with a twist: The fairness judgments were made under cognitive load, which made reasoning close to impossible. This had the predicted result: Without the opportunity to reason, the ratings were identical and showed no hint of hypocrisy.

This experiment is just one illustration of a more general phenomenon. Reasoning is often used to find justifications for performing actions that are otherwise felt to be unfair or immoral (Bandura 1990; Bandura et al. 1996; Bersoff 1999; Crandall & Eshleman 2003; Dana et al. 2007; Diekmann et al. 1997; Haidt 2001; Mazar et al. 2008; Moore et al. 2008; Snyder et al. 1979; for children, see Gummerum et al. 2008). Such uses of reasoning can have dire consequences. Perpetrators of crimes will be tempted to blame the victim or find other excuses to mitigate the effects of violating their moral intuitions (Ryan 1971; for a review, see Hafer & Begue 2005), which can in turn make it easier to commit new crimes (Baumeister 1997). This view of reasoning dovetails with recent theories of moral reasoning that see it mostly as a tool for communication and persuasion (Gibbard 1990; Haidt 2001; Haidt & Bjorklund 2007).

These results raise a problem for the classical view of reasoning. In all these cases, reasoning does not lead to more accurate beliefs about an object, to better estimates of the correctness of one’s answer, or to superior moral judgments. Instead, by looking only for supporting arguments, reasoning strengthens people’s opinions, distorts their estimates, and allows them to get away with violations of their own
moral intuitions. In these cases, epistemic or moral goals are not well served by reasoning. By contrast, argumentative goals are: People are better able to support their positions or to justify their moral judgments.

5. Proactive reasoning in decision making

In the previous section, we have argued that much reasoning is done in anticipation of situations where an opinion might have to be defended, and we have suggested that work on motivated reasoning can be fruitfully reinterpreted in this light. It is not just opinions that may have to be defended: People may also have to put forward arguments to defend their decisions and actions, and they may reason proactively to that end. We want to argue that this is the main role of reasoning in decision making. This claim stands in sharp contrast to the classical view that reasoning about possible options and weighing up their pros and cons is the most reliable way – if not the only reliable way – to arrive at sound decisions (Janis & Mann 1977; Kahneman 2003; Simon 1955). This classical view has in any case been vigorously challenged in much recent research. Some argue that the best decisions are based on intuition and made in split seconds (e.g., see Klein 1998), a view rendered popular by Gladwell (2005). Others maintain that the solution lies with the unconscious and advise us to “sleep on it” (Claxton 1997; Dijksterhuis 2004; Dijksterhuis & van Olden 2006; Dijksterhuis et al. 2006). We briefly review these challenges to the classical view before considering the substantial literature on reason-based choice and interpreting it in the light of the argumentative theory of reasoning.
5.1. To what extent does reasoning help in deciding?

In an initial series of studies, Wilson and his colleagues looked at the effect of reasoning on the consistency between attitudes and behavior (see Wilson & LaFleur 1995; Wilson et al. 1984; 1989a; for review see also Koole et al. 2001; Millar & Tesser 1989; Sengupta & Fitzsimons 2000; 2004; Wilson et al. 1989b). The basic paradigm is as follows: Participants are asked to state their attitude to a given object. In one condition, they have to provide reasons for these attitudes. It has been consistently observed that attitudes based on reasons were much less predictive of future behaviors (and often not predictive at all) than were attitudes stated without recourse to reasons. This lack of correlation between attitude and behavior resulting from too much reasoning can even lead participants to form intransitive preferences (Lee et al. 2008).

Using similar paradigms in which some participants are asked for reasons, it was found that providing reasons led participants to choose items that they were later less satisfied with (Wilson et al. 1993) or that were less in line with the ratings of experts (McMackin & Slovic 2000; Wilson & Schooler 1991). Participants got worse at predicting the results of basketball games (Halberstadt & Levine 1999). People who think too much are also less likely to understand other people’s behavior (Albrechtsen et al. 2009; Ambady & Gray 2002; Ambady et al. 2000). This stream of experiments was later followed up by Dijksterhuis and his colleagues, who introduced a modified paradigm. Here, participants are given lists of features describing different items (such as flats and cars) designed in such a way that some items have more positive features. In the baseline condition, participants had to say which item they
preferred immediately after they had been exposed to these features. In the conscious thought condition, they were left to think about the items for a few minutes. Finally, in the unconscious thought condition, participants spent the same amount of time doing a distraction task. Across several experiments, it was found that the best performance was obtained in this last condition: Unconscious thought was superior to conscious thought (and to immediate decision) (Dijksterhuis 2004; Dijksterhuis & van Olden 2006; Dijksterhuis et al. 2006; 2009).

However, some of Dijksterhuis’s results have proven hard to replicate (Acker 2008; Newell et al., 2009 in press; Thorsteinson & Withrow 2009), and alternative interpretations have been proposed in some cases (Lassiter et al. 2009). In a meta-analysis of this literature, Acker (2008) observed that in only a few experiments was unconscious thought significantly superior to conscious thought, amounting to a null result when all the experiments were taken into account. Even so, there was no significant advantage of conscious thought over immediate choice. This is typically the kind of situation where, according to classical theories, reasoning should help: A new choice has to be made, with the options well delimited and the pros and cons exposed. It is therefore quite striking that reasoning (at least for a few minutes) does not bring any advantage and is sometimes inferior to intuitive, unconscious processes. Finally, studies of decision making in natural environments converge on similar conclusions: Not only are most decisions made intuitively, but when conscious decision-making strategies are used, they often result in poor outcomes (Klein 1998).

In the next subsection, we explore a framework designed to explain such findings by
showing that reasoning pushes people not towards the best decisions but towards decisions that are easier to justify.

5.2. Reason-based choice

Starting in the late 1980s, a group of leading researchers in decision making developed the framework of reason-based choice (for an early review, see Shafir et al. 1993). According to this theory, people often make decisions because they can find reasons to support them. These reasons will not favor the best decisions or decisions that satisfy some criterion of rationality, but decisions that can be easily justified and are less at risk of being criticized. According to the argumentative theory, this is what should happen when people are faced with decisions where they only have weak intuitions. In this case, reasoning can be used to tip the scales in favor of the choice for which reasons are most easily available. One will then at least be able to defend the decision if its outcome proves unsatisfactory.

Reason-based choice is well illustrated in a landmark article by Simonson (1989) in which he studied, in particular, the attraction effect (Huber et al. 1982; for a cross-cultural variation, see Briley et al. 2000). The attraction effect occurs when, given a set of two equally valuable alternatives, a third alternative is added that is just as good as another one of the first alternatives on one trait but inferior on the second trait. This addition tends to increase the rate of choice of the dominating option in a manner not warranted by rational models. Here is one example used in Simonson’s experiments. Participants had to choose between packs of beer that varied along the two dimensions of price and quality. Beer A was of lower quality than beer B but was
also cheaper, and the two attributes balanced in such a way that both beers were regularly chosen in a direct comparison. However, some participants had to choose between these two beers plus beer C, which was more expensive than beer B but not better. When this beer was introduced, participants tended to pick beer B more often. It is easy to account for this finding within the framework of reason-based choice: The poorer alternative makes the choice of the dominating one easy to justify. (“Beer B is of the same quality as but cheaper than this other beer!”) To confirm this intuition, Simonson made and tested the three following predictions: (1) a choice based on reasons should be reinforced when participants have to justify themselves, (2) a choice based on reasons will be perceived as easier to justify and less likely to be criticized, and (3) a choice based on reasons should give rise to more elaborate explanations. The results of three experiments supported these predictions. Moreover, these results also showed that participants who made choices based on reasons tended to make choices that fitted less well with their own preferences as stated before the choice was made. Finally, another set of experiments demonstrated that, when participants were able to use their intuitions more, because they were familiar with the alternatives or because the descriptions of these alternatives were more detailed, they were less prone to the attraction effect (Ratneshwar et al. 1987). Several well-known challenges to the view of humans as making rational decisions thanks to their reasoning abilities have been, or can be, reinterpreted as cases of reason-based choice.

5.3. What reason-based choice can explain
5.3.1. Disjunction effect. The *sure-thing principle* (Savage 1954) states that, when someone favors A over B if event E happens and keeps the same preference ordering if E does not happen, then her choices should not be influenced by any uncertainty about the occurrence of E. Shafir and Tversky (1992; Tversky & Shafir 1992) have recorded several violations of this principle. For instance, we can compare the reaction of participants to the following problems (Tversky & Shafir 1992, p. 306):

*Win/lose versions*

Imagine that you have just played a game of chance that gave you a 50% chance to win $200 and a 50% chance to lose $100. The coin was tossed and you have either won $200 or lost $100. You are now offered a second identical gamble: 50% chance to win $200 and 50% chance to lose $100. Would you?: (a) accept the second gamble. (b) reject the second gamble.

Whether they have won or lost in the first gamble, a majority of participants accept the second gamble. However, they are likely to do so for different reasons: In the win scenario, they reason that they can easily risk losing half of the $200 they have just won. In the lose scenario, however, they might take the second gamble as an opportunity to make up for their previous loss. In these two cases, while the choice is the same, the reasons for making it are incompatible. Thus, when participants do not know what is going to be the outcome of the first bet, they have more trouble justifying the decision to accept the second gamble: The reasons seem to contradict each other. As a result, a majority of participants who do not know the result of the
first gamble reject the second gamble even though they would have accepted it whatever the result of the first gamble. The authors tested this explanation further by devising a comparison that had the same properties as the one just described, except that the reasons for making the “accept” decision were the same irrespective of the outcome of the first gamble. In this case, participants made exactly the same choices whether or not they knew the result of the first gamble (for a similar experiment with a variant of the prisoner’s dilemma, see Croson 1999).

5.3.2. Sunk-cost fallacy. The sunk-cost fallacy is the “greater tendency to continue an endeavor once an investment in money, effort, or time has been made” (Arkes & Blumer 1985, p. 124). A well-known real-life example is that of the Concorde: The British and French governments decided to keep paying for a plane that they knew would never turn a profit. Arkes and Ayton (1999) have argued that such mistakes result from an unsatisfactory use of explicit reasons such as “do not waste.” We will briefly review the evidence they presented, and add more.

First of all, Arkes and Ayton (1999) contrast the robust sunk-cost effects observed in humans (Arkes & Blumer 1985; Garland 1990; Staw 1981) with the absence of such mistakes among animals. They also point out that children do not seem prone to this error (for more recent, convergent evidence, see Klaczynski & Cottrell 2004; Morsanyi & Handley 2008). If reasoning were not the cause of this phenomenon but the cure for it, the opposite would be expected. Finally, some experiments have varied the availability of justifications – a factor that should not be relevant for standard models of decision making. Thus, when participants can justify
the waste, they are less likely to be trapped by sunk costs (Soman & Cheema 2001). By contrast, when participants find it harder to justify changing their course of actions, they are more likely to commit the fallacy (Bragger et al. 1998; 2003).

5.3.3. Framing. Framing effects occur when people give different answers to structurally similar problems depending on their wording – their “frame” (Tversky & Kahneman 1981). Our intuitions are generally blamed for these effects (Kahneman 2003). Another explanation that can be seen as either complementary or alternative to this one is that different frames make some reasons more or less available, thus modifying the way reasoning affects our decisions. Several results support this interpretation (see McKenzie 2004; McKenzie & Nelson 2003). First, as mentioned earlier, participants who reason more about the tasks are more influenced by framing effects (Igou & Bless 2007). Second, when groups make decisions on framed problems, the groups tend to converge on the answer that is supported by the strongest reasons (McGuire et al. 1987; Milch et al. 2009; Paese et al. 1993). If the participants’ answers were truly based on their intuitions, the answer proposed by the group would tend to be the mean of these different intuitions (Allport 1924; Farnsworth & Behner 1931). Instead, these findings have to be explained within the framework of the Persuasive Argument Theory (Vinokur 1971; Vinokur & Burnstein 1978), showing that the decisions are based on reasons.

5.3.4. Preference inversion. The ability to evaluate preferences correctly is necessary for economic models of decision making, but preferences can vary dramatically
depending on the way they are measured. Someone may rate A higher than B and still choose B over A (Bazerman et al. 1992; Irwin et al. 1993; Kahneman & Ritov 1994; Slovic 1975; Tversky et al. 1988). For instance, the relative rating of two objects can vary or even be reversed, depending on whether they are rated separately or jointly (Hsee 1996b; 1998; Hsee et al. 1999). Thus, when the following two objects are presented in isolation – a music dictionary with 10,000 entries that is “like new,” and one with 20,000 entries and a torn cover – people rate the one with 10,000 entries more highly. However, when people have to choose between the two, they favor the one that has more entries, despite the torn cover (Hsee 1996b). Such effects fit perfectly in the current framework: People choose an alternative because they can provide “a compelling argument for choice that can be used to justify the decision to oneself as well as to others” (Tversky et al. 1988, p. 372). In the foregoing example, people lack reliable intuitions – they cannot tell how many entries a good music dictionary should have. Lacking such intuitions, they fall back on reasoning and let their judgments be guided by ease of justification – in this case, the condition of the dictionary that easily justifies a high or low price. On the other hand, dimensions with numerical values will often provide compelling justifications when options are presented jointly. This bias can lead to suboptimal decisions (Hsee & Zhang 2004).

More generally, “decision-makers have a tendency to resist affective influence, and to rely on rationalistic attributes to make their decisions” (Hsee et al. 2003, p. 16; see also Okada 2005). Indeed, rationalistic attributes make for easy justifications. For instance, in one experiment, participants had either to choose between the following two options or to rate them: A roach-shaped chocolate
weighing 2 ounces and worth 2 dollars, and a heart-shaped chocolate weighing half an ounce and worth 50 cents (Hsee 1999). A majority (68%) of participants chose the roach-shaped chocolate, even though more than half (54%) thought they would enjoy the other more. The participants who chose the bigger, roach-shaped chocolate did so because the feeling of disgust, being “irrational,” was hard to justify, especially compared to the difference in price and size. However, in light of the results from the psychology of disgust (e.g., Rozin et al. 1986), we can tell that their choice was certainly the wrong one.

5.3.5. Other inappropriate uses of reasons. Many other inappropriate uses of reasons have been empirically demonstrated. Investors’ decisions are guided by reasons that seem good but are unrelated to real performance (Barber et al. 2003). People will use a rule such as “more variety is better” or “don’t pick the same things as others” to guide their decisions, even when less variety or more conformity would actually be more in line with their preferences (Ariely & Levav 2000; Berger & Heath 2007; Simonson 1990). Use of a rule such as “don’t pay for delays” will lead to behaviors that go against one’s own interest (Amir & Ariely 2003). When forecasting their affective states, people rely on explicit lay theories (Igou 2004), which will often lead them astray (Hsee & Hastie 2006). Because “it’s better to keep options open,” people will be reluctant to make an unalterable decision even when they would be better off making it (Gilbert & Ebert 2002). When indulging in a hedonic act, people feel they need a reason for such indulgence, even though this does not actually change the quality of the experience (Xu & Schwarz 2009). Reason-based choice has
also been used to explain effects related to loss aversion (Simonson & Nowlis 2000),
the effect of attribute balance (Chernev 2005), the tendency to be overwhelmed by
too much choice (Scheibehenne et al. 2009; Sela et al., 2009), the feature creep effect
(Thompson et al. 2005a), the endowment effect (Johnson et al. 2007), aspects of time
discounting (Weber et al. 2007), and several other departures from the norms of
rationality (Shafir et al. 1993).

Another sign that reason-based choice can lead to nonnormative outcomes is
that sometimes reasons that are not relevant to the decision will nonetheless play a
role. For instance, the same irrelevant attribute will sometimes be used as a reason for
choosing an item (Carpenter et al. 1994) and sometimes as a reason for rejecting it
(Simonson et al. 1993; 1994), depending on what decision it makes easier to justify
(Brown & Carpenter 2000). People will also be influenced by irrelevant pieces of
information because they find it hard to justify ignoring them (Tetlock & Boettger

All of these experiments demonstrate cognitively unsound uses of reasoning.
There are two ways to explain these findings. One could argue that these are instances
of a mechanism designed for individual cognition, and in particular for decision
making, that sometimes gets misused. According to the argumentative theory,
however, the function of reasoning is primarily social: In particular, it allows people
to anticipate the need to justify their decisions to others. This predicts that the use of
reasoning in decision making should increase the more likely one is to have to justify
oneself. This prediction has been borne out by experiments showing that people will
rely more on reasons when they know that their decisions will later be made public
(Thompson & Norton 2008) or when they are giving advice (in which case one has to be able to justify oneself [see Kray & Gonzalez 1999]). By contrast, when they are choosing for others rather than for themselves, they are less prone to these effects because there is then less need for a utilitarian, justifiable decision (Hamilton & Thompson 2007). Finally, it should be stressed that the picture of reasoning painted in these studies may be overly bleak: Demonstrations that reasoning leads to errors are much more publishable than reports of its successes (Christensen-Szalanski & Beach 1984). Indeed, in most cases, reasoning is likely to drive us towards good decisions. This, we would suggest, is mostly because better decisions tend to be easier to justify. The reasons we use to justify our decisions have often been transmitted culturally and are likely to point in the right direction – as when people justify their avoidance of sunk-cost mistakes by using the rule they have learned in class (Simonson & Nye 1992). In such cases, the predictions of the argumentative theory coincide with those of more classical theories. However, what the results just reviewed show is that, when a more easily justifiable decision is not a good one, reasoning still drives us in the direction of ease of justification. Even if they are rare, such cases are crucial to comparing the present theory (reasoning drives us to justifiable decisions) with more classical ones (reasoning drives us to good decisions).

6. Conclusion: Reasoning and rationality

Reasoning contributes to the effectiveness and reliability of communication by enabling communicators to argue for their claim and by enabling addressees to assess
these arguments. It thus increases both in quantity and in epistemic quality the information humans are able to share.

We view the evolution of reasoning as linked to that of human communication. Reasoning, we have argued, enables communicators to produce arguments to convince addressees who would not accept what they say on trust; it enables addressees to evaluate the soundness of these arguments and to accept valuable information that they would be suspicious of otherwise. Thus, thanks to reasoning, human communication is made more reliable and more potent. From the hypothesis that the main function of reasoning is argumentative, we derived a number of predictions that, we tried to show, are confirmed by existing evidence. True, most of these predictions can be derived from other theories. We would argue, however, that the argumentative hypothesis provides a more principled explanation of the empirical evidence (in the case of the confirmation bias, for instance). In our discussion of motivated reasoning and of reason-based choice, not only did we converge in our prediction with existing theories, we also extensively borrowed from them. Even in these cases, however, we would argue that our approach has the distinctive advantage of providing clear answers to the why-questions: Why do humans have a confirmation bias? Why do they engage in motivated reasoning? Why do they base their decisions on the availability of justificatory reasons? Moreover, the argumentative theory of reasoning offers a unique integrative perspective: It explains wide swaths of the psychological literature within a single overarching framework.

Some of the evidence reviewed here shows not only that reasoning falls short of delivering rational beliefs and rational decisions reliably, but also that, in a variety
of cases, it may even be detrimental to rationality. Reasoning can lead to poor outcomes not because humans are bad at it but because they systematically look for arguments to justify their beliefs or their actions. The argumentative theory, however, puts such well-known demonstrations of “irrationality” in a novel perspective. Human reasoning is not a profoundly flawed general mechanism; it is a remarkably efficient specialized device adapted to a certain type of social and cognitive interaction at which it excels.

Even from a strictly epistemic perspective the argumentative theory of reasoning does not paint a wholly disheartening picture. It maintains that there is an asymmetry between the *production* of arguments, which involves an intrinsic bias in favor of the opinions or decisions of the arguer whether or not they are sound, and the *evaluation* of arguments, which aims at distinguishing good arguments from bad ones and hence genuine information from misinformation. This asymmetry is often obscured in a debate situation (or in a situation where a debate is anticipated). People who have an opinion to defend don’t really evaluate the arguments of their interlocutors in a search for genuine information but rather consider them from the start as counterarguments to be rebutted. Still, as shown by the evidence reviewed in section 2, people are good at assessing arguments and are quite able to do so in an unbiased way, provided they have no particular axe to grind. In group reasoning experiments where participants share an interest in discovering the right answer, it has been shown that *truth wins* (Laughlin & Ellis 1986; Moshman & Geil 1998). While participants in collective experimental tasks typically produce arguments in favor of a variety of hypotheses, most or even all of which are false, they concur in
recognizing sound arguments. Since these tasks have a demonstrably valid solution, truth does indeed win. If we generalize to problems that do not have a provable solution, we should at least expect good arguments to win, even if this is not always sufficient for truth to win (and, in section 2, we have reviewed evidence that this is indeed the case). This may sound trivial, but it is not. It demonstrates that, contrary to common bleak assessments of human reasoning abilities, people are quite capable of reasoning in an unbiased manner, at least when they are evaluating arguments rather than producing them, and when they are after the truth rather than trying to win a debate.

Couldn’t the same type of situation that favors sound evaluation favor comparable soundness in the production of arguments? Note, first, that situations where a shared interest in truth leads participants in a group task to evaluate arguments correctly are not enough to make them produce correct arguments. In these group tasks, individual participants come up with and propose to the group the same inappropriate answers that they come up with in individual testing. The group success is due to, first and foremost, the filtering of a variety of solutions, achieved through evaluation. When different answers are initially proposed and all of them are incorrect, then all of them are likely to be rejected, and wholly or partly new hypotheses are likely to be proposed and filtered in turn, thus explaining how groups may do better than any of their individual members.

Individuals thinking on their own without benefiting from the input of others can assess only their own hypotheses, but in doing so, they are both judge and party, or rather judge and advocate, and this is not an optimal stance for pursuing the truth.
Wouldn’t it be possible, in principle, for an individual to decide to generate a variety of hypotheses in answer to some question and then evaluate them one by one, on the model of Sherlock Holmes? What makes Holmes such a fascinating character is precisely his preternatural turn of mind operating in a world rigged by Conan Doyle, where what should be inductive problems in fact have deductive solutions. More realistically, individuals may develop some limited ability to distance themselves from their own opinion, to consider alternatives and thereby become more objective. Presumably this is what the 10% or so of people who pass the standard Wason selection task do. But this is an acquired skill and involves exercising some imperfect control over a natural disposition that spontaneously pulls in a different direction.

Here, one might be tempted to point out that, after all, reasoning is responsible for some of the greatest achievements of human thought in the epistemic and moral domains. This is undeniably true, but the achievements involved are all collective and result from interactions over many generations (on the importance of social interactions for creativity, including scientific creativity, see Csikszentmihalyi & Sawyer 1995; Dunbar 1997; John-Steiner 2000; Okada & Simon 1997). The whole scientific enterprise has always been structured around groups, from the Lincean Academy down to the Large Hadron Collider. In the moral domain, moral achievements such as the abolition of slavery are the outcome of intense public arguments. We have pointed out that, in group settings, reasoning biases can become a positive force and contribute to a kind of division of cognitive labor. Still, to excel in such groups, it may be necessary to anticipate how one’s own arguments might be evaluated by others and to adjust these arguments accordingly. Showing one’s ability
to anticipate objections may be a valuable culturally acquired skill, as in medieval
disputationes (see Novaes 2005). By anticipating objections, one may even be able to
recognize flaws in one’s own hypotheses and go on to revise them. We have
suggested that this depends on a painstakingly acquired ability to exert some limited
control over one’s own biases. Even among scientists, this ability may be uncommon,
but those who have it may have a great influence on the development of scientific
ideas. It would be a mistake, however, to treat their highly visible, almost freakish,
contributions as paradigmatic examples of human reasoning. In most discussions,
rather than looking for flaws in our own arguments, it is easier to let the other person
find them and only then adjust our arguments, if necessary.

In general, one should be cautious about using the striking accomplishments
of reasoning as proof of its overall efficiency, since its failures are often much less
visible (see Ormerod 2005; Taleb 2007). Epistemic success may depend to a
significant extent on what philosophers have dubbed epistemic luck (Pritchard 2005);
that is, chance factors that happen to put one on the right track. When one happens to
be on the right track and “more right” than one could initially have guessed, some of
the distorting effects of motivated reasoning and polarization may turn into blessings.
For instance, motivated reasoning may have pushed Darwin to focus obsessively on
the idea of natural selection and explore all possible supporting arguments and
consequences. But, for one Darwin, how many Paleys?

To conclude, we note that the argumentative theory of reasoning should be
congenial to those of us who enjoy spending endless hours debating ideas – but this,
of course, is not an argument for (or against) the theory.
Acknowledgements

We are grateful to Paul Bloom, Ruth Byrne, Peter Carruthers, Nick Chater, Jon Haidt, Ira Noveck, Guy Politzer, Jean-Baptiste Van der Henst, Deirdre Wilson, and four anonymous reviewers for useful suggestions and criticisms on earlier versions of this article. Our work has been supported by a PhD grant of the DGA (Paris) to Hugo Mercier and by the CSMN (Oslo).

NOTES

1. Recently, reasoning has been used simply as a synonym of inference and is then unproblematically attributed to infants (Spelke & Kinzler 2007) or to nonhuman animals (Blaisdell et al. 2006). In this article, however, we use “reasoning” in its more common and narrower sense. The content of the article should make it clear why we see this as a principled terminological choice.

2. Our functional hypothesis will be tested without reference to specific mechanisms (as is common in evolutionary biology). Even if one can ask to what extent attributing an argumentative function to reasoning suggests or favors a specific algorithmic account, this will not be the focus of this article. There is, in any case, no obvious clash between our functional account and various algorithmic accounts that have been offered, for instance, by Evans (2007), Johnson-Laird (2006), or Rips (1994).
3. In the psychology of reasoning, some tasks can be described as *production tasks* because participants have to produce a logically valid conclusion from a set of premises. However, these tasks are very different from the production of arguments in a debate. In a dialogic context, one starts from the conclusion and tries to find premises that will convince one’s interlocutor. It is this meaning of *production* that is relevant here.

4. It should be noted that this spotty record may be partly explained by very artificial conditions: In the vast majority of group experiments, participants are asked to interact with people they don’t know and will never meet again, and to perform tasks that have no bearing on their lives outside the laboratory. When any of these factors is made more natural, performance improves. Debates about political matters between laypeople often lead to epistemic improvement (Landemore, in press; Mercier & Landemore, submitted). Groups that are used to working together are much more efficient (Michaelsen et al. 1989). And collaborative learning is hugely successful in schools (Slavin 1995).


6. Incidentally, another advantage of the theory suggested here is that it makes testable predictions about the contexts that should motivate the use of reasoning;
namely, contexts in which real or anticipated argumentation takes place. This contrasts with standard dual-process theories, which do not have a principled and testable way of predicting when system 2 reasoning should be triggered.

7. It may be worth mentioning that what general motivation fails to bring about is efficient or unbiased reasoning rather than reasoning per se. If you pay people to get the right answer in, say, the Wason selection task, they may reason more but will still be as biased, and their answer will still be wrong.

8. The Delphi technique is a method of forecasting that can be seen as trying to make the best of the confirmation bias by having different experts critique on another’s predictions and justify their own predictions. Its effectiveness shows that, in an appropriate context, the confirmation bias can be conducive to very good performance (Green et al. 2007; Keeney et al. 2001; Powell 2003; Rowe & Wright 1999; Tichy 2004).

9. Note that motivated, or motivation, as used here do not refer to conscious motivation based on reasons, as in “I’m going to think of arguments supporting this opinion of mine in case someone questions me later.” Instead, it refers to processes that influence either the direction or the triggering of reasoning in a mostly unconscious manner. Even though a lawyer, for instance, can consciously trigger reasoning and influence its direction, this is the exception and not the rule. Generally, people (including lawyers) have limited control over the triggering of reasoning or the direction it takes.

10. Attitude polarization is most likely to occur in individuals who hold a very strong attitude with a high degree of confidence. The problem is, then, that these
individuals will tend to fall at one end of the attitude scale before reading the arguments, which makes it close to impossible to detect any movement towards a more extreme attitude. This can explain, at least in part, the failed replications by Kuhn and Lao (1996) and Miller et al. (1993).

11. Incidentally, this does not explain all forms of belief perseverance: Other mechanisms may be involved in some instances (e.g., see Anderson et al. 1980), but the availability of arguments supporting the discredited belief may still be crucial (see Anderson et al. 1985).

12. It has been shown that pigeons fall prey to the fallacy but only when no indication was given that they were in such a situation (Navarro & Fantino 2005). The instructions received by human participants always make this point clear, so these experiments confirm the point made by Arkes and Ayton (1999).

References [Mercier & Sperber] [aHM]


Allport, F. (1924) *Social psychology*. Houghton Mifflin. [aHM]


assessments of ability. *Journal of Personality and Social Psychology* 57(6):1082–90. [aHM]


decision to escalate commitment to an ongoing project. *Journal of Applied
Psychology* 75(6):728–31. [aHM]


D. Griffin & D. Kahneman, pp. 167–84. Cambridge University Press. [aHM]

forecasting of changeable outcomes. *Journal of Personality and Social
Psychology* 82(4):503–14. [aHM]


reasoners or pragmatic virtuosos? Relevance and the deontic selection task.
*Cognition* 81(2):69–76. [aHM]

[aHM]

groups: Delphi and prediction markets compared. Foresight: The International
Journal of Applied Forecasting Fall issue: 17–21. [aHM]


Mercier, H. (submitted a) The argumentative function of reasoning: Evidence from developmental and educational psychology. [aHM]

Mercier, H. & Landemore, H. (submitted) Reasoning is for arguing: Understanding the successes and failures of deliberation. [aHM]


Ormerod, P. (2005) *Why most things fail: Evolution, extinction and economics*. Faber & Faber. [aHM]


Ryan, W. (1971) *Blaming the victim*. Pantheon. [aHM]


Steiner, I. D. (1972) *Group processes and productivity*. Academic Press. [aHM]


Tetlock, P. E. (1998) Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right. *Journal of Personality and Social Psychology* 75(3):639–52. [aHM]


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. [aHM]


