DRAFT – NOT TO BE QUOTED

Our pigheaded core:

How we became smarter to be influenced by other people.

Hugo Mercier

Philosophy, Politics and Economics Program

University of Pennsylvania

313 Cohen Hall

249 South 36th Street

Philadelphia, PA 19104

hmercier@sas.upenn.edu

http://hugo.mercier.googlepages.com/

To be published in *Signaling, Commitment, and Emotion*,

Edited by Brett Calcott, Richard Joyce, & Kim Sterelny (MIT Press)

Word count (main text): 8,721

(Other) people's gullibility is a common source of complaint in the political world. Republicans lament that Democrats naively trust the 'liberal media'. Democrats wonder how Republicans can be so credulous as to believe Fox News. In this kind of attack, gullibility is often equated with lack of sophistication, the subtext is "How can they be so stupid?" ("Sophisticated" is a common antonym of "gullible"). Indeed, there seems to be a widespread intuition that the best way to influence people is to stop them from thinking. Politicians, newscasters and educators are wont to dumb down their messages; ad men try to distract us so that their slogans will remain unexamined; interrogators try to break suspects' ability to reason through continuous questioning or sleep deprivation (or worse). Yet it is possible to argue that this intuition is profoundly misguided and that, overall, the best way to influence people is to tap into their most sophisticated psychological mechanisms, especially the complex calibration of trust and reasoning. In support of this claim, I will first offer an evolutionary argument and then proceed to "peel off" the mechanisms we use to evaluate information to reveal older mechanisms that are harder to influence: our pigheaded core.

The joys and dangers of communication

All the important decisions we make in our life will have been profoundly affected by communicated information: who we will befriend, who will be our partner(s), where we will live, what career we will follow. It would be difficult to deny that our ability to communicate, and all the things which it enables, from collective action to cultural transmission, have played a major role—probably *the* major role—in the evolutionary success of our species. Underlying the importance of communication is its amazing efficiency: on a first approximation, language is very close to telepathy. However, great dangers accompany such a powerful mean of influence. Given the uncanny ability of other individuals to create representations in our minds, and the power these representations can have in shaping our decisions, there is an ever present possibility of abuse. People can manipulate each other:

senders can lie, deceive, and more generally take advantage of listeners. Even if one were to discount the possibility of purely Machiavellian motives, a simple lack of consideration of other people's interests would lead to widespread trouble. (One could truthfully state that "I want you to want to give me $10,000"—it would still be a bad idea to comply.)

Dawkins and Krebs were among the first to draw attention to the evolutionary problems raised by communication (Dawkins & Krebs, 1978; Krebs & Dawkins, 1984). For communication to be evolutionarily stable, it has to benefit both senders and receivers. If senders do not benefit from communication they stop sending (they become "mute"), and if it is the receivers who lack benefits, then they stop receiving (they become "deaf"). That communication often breaks down in zero sum games, even in an intensely communicative species as ours, is immediately apparent from the behavior of, for instance, poker players: the "poker face" is an extreme example of what happens when there is no incentive whatsoever to communicate.

Once senders can influence receivers, they will often have incentives to take advantage of them. Alarm calls could be used to distance others from a valuable resource, food calls could be use to lure other individuals to a given place, etc. This means that in a great many cases the interests of the receivers—and, therefore, the stability of communication—will have to be secured by some additional mechanism. Many such devices have been discovered across the animal world. For instance, if a signal is costly it can be used as a reliable indicator of the capacity and motivation of the sender to pay these costs (Zahavi & Zahavi, 1997). The incentives of senders and receivers can also be more easily aligned when their genetic interests overlap—although examples such as conflicts between mothers and their fetuses dramatically attest that this is far from being a foolproof solution (Haig, 1993). The stability of human communication cannot easily be explained through these means: we routinely communicate with non-genetically related individuals and the bulk of our communication is cheap talk (for economists, Farrell & Rabin, 1996) or low-cost signaling (for biologists, Maynard-Smith, 1994). Another solution is to take advantage of the specific characteristics of a communication domain (Sterelny, in press). The structure of the communicative environment—such as redundancies between different sources—can provide a cheap way to evaluate the validity of a piece of information. Some contents can also be easily verified—demonstrations of skills, for instance, are hard to fake. When communication only bears on a limited domain, dedicated mechanisms relying on the specifics of the situation can evolve. In humans, this could be the case for some forms of emotional contagion, for instance (see note 4 below). But human communication differs from anything else that exists in the animal kingdom because we can talk about practically anything. This means that other mechanisms, that do not rely on the specifics of a given domain, are likely to have evolved in order to maintain the stability of this very general form of communication. Receivers can *filter* communicated information; they can use a variety of means to determine which communicated information they should pay heed to.

Trust calibration may be the more obvious of these mechanisms: when we are told something, we will weight it differently depending on its source—a trusted doctor or a quack, a friend or a stranger, etc. But, given the huge influence of communication on our lives, calibration of trust can only be one of the many devices we use to evaluate communicated information. Some of these mechanisms will deal mostly with the ‘compliance’ level: should I act on the basis of communicated information? Others will be dedicated to epistemic vigilance: maximizing the chances that the beliefs we accept are true (Mascaro & Sperber, 2009; Sperber et al., in press).

If we assume that, like other complex adaptations (Pinker & Bloom, 1990), human communication—consisting mostly of, but not restricted to, language—evolved relatively gradually since our last common ancestors with the chimpanzees, becoming increasingly efficient and influential, it follows that the filtering mechanisms must also have evolved gradually, in line with communication's ever growing importance in our ancestors' lives. There are two ways for such filtering mechanisms to evolve, each of which may have played a role at different points of our evolution. Filtering mechanisms can evolve in order to *reject more and more information*. Let's imagine that at some stage of evolution individuals are not very discriminating and that as a result they *accept too much information*. Since receivers are not very discriminating, this might leave a window for senders to evolve towards more skillful manipulation. Receivers are then under pressure to evolve better filtering mechanisms—otherwise they stop benefitting from communication. This can be dubbed the "Machiavellian" view. But filtering mechanisms could also have evolved in order to *accept more and more information*. Here we start from a stage in which receivers *reject too much information*. If, then, communication becomes more important, receivers are under pressure to evolve better filtering mechanisms so that they can accept information they were previously rejecting.[1,2]

Which of these two processes played the major role in the evolution of human filtering mechanisms? Intuitively, it might seem that erring on the side of caution should be the most effective strategy. However, making a strong a priori argument to that effect would require estimates of the costs and benefits of rejecting too much true information and of accepting too much false information that may not be within our reach. Moreover, we cannot examine directly the psychological mechanisms of our ancestors. At best, we can hope to make

[1] The increase in acceptance of communicated information is to be understood in relative terms: the ratio of information accepted to information rejected. However, given that the total amount of information should also increase (this is the point of the evolution of more sophisticated filtering mechanisms in the first place), it is possible that the absolute amount of information rejected will also increase.

[2] Even though this chapter will not dwell on them, in tandem with these new filtering mechanisms, senders should evolve new mechanisms to help receivers filter information more efficiently, so that senders can exert more influence on receivers through an increase in accepted information. Moreover, given that senders are the initiators of communication, they are likely to be its greatest beneficiary, and so we should even expect them to bear most of the costs of the mechanisms that will make communication more efficient.

informed inferences regarding their communicative system, but that would not get us very far in our present endeavor. However, we may be able to find remnants of earlier filtering mechanisms in modern humans. Much in the same way as linguists study ‘fossils’ in modern languages to make inferences regarding previous stages of language evolution (e.g. Jackendoff, 1999), it might be possible to observe older filtering mechanisms still at play in modern humans. Accordingly, the present chapter will review evidence related to these older filtering mechanisms and try to show that they tend to err on the side of caution—that they reject too much information rather than not enough—and that new mechanisms evolved to make us accept more information.

For the ‘fossil’ approach to work, however, these older mechanisms need to have been preserved. More specifically, the new mechanisms would have become fine-grained regulators of these older mechanisms. This entails a specific, highly modular view of the mind. Another perspective could also be envisioned, one in which the older mechanisms are ‘cannibalized’: they are modified and put to new uses. Here again, it might be possible to make an a priori argument in favor of the former view. According to this former view, the main role played by the new mechanisms would be to override the negative verdicts of the older mechanisms when they deem a piece of communicated information to be, in fact, beneficial. With such a design, if, for some reason, the new mechanisms are unable to function properly, then the older mechanisms will take over and protect the individual from the costs of accepting too much information. General principles of design—which apply from cells to complex artifacts—would seem to favor this solution as it is more modular, more robust, and it relies more on regulation (Carlson & Doyle, 2002; Kitano, 2004; Wagner, 2005). However, as is the case for the previous argument, these a priori reasons are unlikely to sway most readers. So I will rely on the evidence reviewed above to try and tip the scales in favor of the hypothesis that older filtering mechanisms are still present but that they have been increasingly regulated by layers of more recent filtering mechanisms.

In what follows, I will try to show (i) *that older, less sophisticated mechanisms filtering communicated information are still present in humans* and (ii) *that these mechanisms are overly cautious and reject more information than more recent mechanisms*. Layers of filtering mechanisms will be peeled off to reveal or 'pigheaded core'. To do so, I will have to find situations, experimental or otherwise, in which the workings of the more recent layers were perturbed, and the action of the older mechanism exposed, starting with the most extreme case: that of subliminally presented information.

The core: competition between goals

Humans are endowed with a wealth of dedicated mechanisms that filter incoming communicated information (Sperber et al., in press). What is left is we try to remove all these dedicated mechanisms? What is left is a very simple mechanism that, even if it is not its main purpose, can already protect us against communicated information—even subliminally communicated information: competition between goals

In cognitively complex species, many goals or plans are competing at any given point for control of our motor system (Sperber, 2005). Stimuli perceived in the environment will compete with each other as well as with previous plans in such a manner that there will nearly never be a mandatory behavioral reaction to any type of input. This picture is controversial, but a review of the literature is out of question here, so I will only offer an example rendered convincing by the counter-intuitiveness of its conclusion.

Neurobiologists have devoted a lot of attention to the tail-flip escape response of the crayfish, to such a point that it has become "one of the best-understood neural circuits in the animal kingdom" (Edwards, Heitler, & Krasne, 1999, 153, to which this section is heavily indebted). At first sight, this might seem to be a poster-boy for the reflex; a single neuron commands a single behavior that is ecologically crucial: a flight response. But careful research has shown that the action of this “reflex” is in fact modulated by many factors. If escape would be blocked by physical constraints, then it stops being triggered, plausibly to prevent the animal from hurting itself. When the crayfish is feeding, the perception of threat has to be more serious or more persistent to trigger escape: this makes sense given that there is a higher cost associated with the departure from the food resources (Krasne & Lee, 1988). Likewise, if the crustacean is already involved in an incompatible behavior—such as backward walking or the defense posture—then the tail-flip escape response is strongly inhibited (Beall, Langley, & Edwards, 1990). Finally, and even more impressively, the social status of the crayfish will heavily influence its escape response: if the individual is a subordinate, it will moderate its behavior and switch to (even more) “flexible, nonreflex (‘voluntary’) types of escape” (Krasne, Shamsian, & Kulkarni, 1997, 709). Dominants, however, seem to ignore the presence of subordinates and go about their behavior unperturbed (in a manner paralleled by human dominants, see, e.g. (Fiske, 1993). All of these things are performed routinely by crayfish. While this is far from a demonstration, it shows at least that any assumption of “reflexiveness” should be very seriously tested before it is accepted and that, in the meantime, it seems reasonable to assume that the vast majority—probably all—of our behaviors are the result of some kind of modulation, including competition among different goals.

Mechanisms designed to regulate motor control in such a way are very old and have clearly not evolved specifically to deal with problems related to filtering communicated information. Still, their action should play an important role in this domain because goals generated or influenced by communicated information—as any other goal— have to win this competition to cause any kind of behavioral effect. According to the present argument, when it comes to communicated information, this mechanism should be expected to have a very stringent baseline. Our behavior should only be minimally influenced by communicated information when we are deprived of any means of evaluation. But it is not easy to test for such a baseline because in the vast majority of cases we have at least one means to evaluate the information, namely its source. Even when the source is someone we know very little about, or is anonymous, we can always venture educated guesses about what kind of individual she is and how much trust we can grant her. There is one case, however, in which we seem to be deprived of any mean to evaluate a piece of communicated information: subliminal influence.

In subliminal influence people are exposed to stimuli that lie completely outside of their awareness. The most famous cases are words flashed on a screen too quickly for people to consciously perceive them. Since the 1950s, subliminal influence has been a recurrent source of fears ("I'm being constantly manipulated!") and hopes ("The tapes that I listen while I sleep will make me smart and self-confident"). However, what people thought were conclusive results proved to be the fraudulent invention of a disgruntled ad man (Weir, 1984) and the subsequent forty years of research would fail to detect any sign of subliminal influence (see for instance Greenwald, Spangenberg, Pratkanis, & Eskenazi, 1991; Moore, 1982; Pratkanis & Aronson, 1992). More recently, some results have started to surface showing reliable effects of subliminal stimuli. The big difference introduced by these new experiments is that the stimuli are coherent with the previous goals of the participants. For instance, thirsty participants flashed with subliminal words related to thirst drink more water than participants flashed with neutral words. The exact same stimuli, however, have no effect whatsoever on non-thirsty participants (Strahan, Spencer, & Zanna, 2002, see also Berridge & Winkiehnan, 2003; Dijksterhuis & Bargh, 2001). As Bargh puts it: "The main reason for the recent success is that researchers are taking the consumer's (experimental participant's) current goals and needs into account." (Bargh, 2002, 282-83). So, while subliminal influence may make you drink a little more water if you are thirsty (thanks to 50 years of intensive research), good old influence can make most of us inflict enough electric shocks to risk killing a fellow human being (Milgram, 1974). Given the power many people attributed to subliminal influence, this outcome is rather ironic and confirms the idea that the older filtering mechanisms are not very responsive to communicated information.

Ostensive and non-ostensive communication

We have seen that when filtering mechanisms are stripped to their bare minimum—when stimuli are perceived subliminally—they make it very hard for people to be influenced: only small changes in behaviors that were already planned are allowed. This is coherent with the hypothesis that core mechanisms are still present today that protect us by severely diminishing the influence of communicated information. However, there are many layers between this core and the most recent filtering mechanisms. In order to keep peeling off the layers of filtering mechanisms, it is now useful to introduce the distinction between ostensive and non-ostensive communication.

Here we will define an ostensive stimulus as one that aims at attracting an audience's attention (Sperber & Wilson, 1995). The bulk of human communication is ostensive-inferential: senders ostensively provide evidence that will help the receiver infer the sender's meaning (Sperber & Wilson, 1995). However, other channels of communication are non ostensive. For instance, if something frightens me, my feelings are likely to be reflected in a facial expression. This facial expression is communicative, but it is not ostensive: I do not intend to attract your attention to the fact that my face is harboring this expression[3]. By

[3] In the literature on the evolution of communication, there is an important distinction between signal, cues and coercion. For something to be a signal mechanisms designed to send it and mechanisms designed to receive it must have evolved specifically for that purpose. By

contrast, I could mime the expression of fear in an ostensive manner—I could start by making eye contact. In this case, I'm trying to ostensively communicate something to you—for instance, that I think that something frightening is going to happen in the movie we are watching together.

There are several differences between ostensive and non-ostensive communication, but the most relevant to the present endeavor is that ostensive communication provides receivers with an additional layer of protection against communicated information. The reason is that when a receiver infers a sender's meaning, the output of this inference is in a metarepresentational format: "sender means P", which in the case of a standard assertion, entails "sender intends me to believe P". It makes a lot of sense, from a filtering point of view, for the "P" not to be automatically disembedded from its metarepresentational context (see (Sperber, 1997). While it is so embedded, it is mostly harmless: thinking that you intend me to believe P when I doubt P or even have reasons to believe that it is a lie is not only innocuous, but it can be highly informative (I will trust you less in the future; I can try to understand your motive for lying to me).

Even though the lack of automatic disembedding makes sense from the present point of view, it is not generally accepted. In particular, one line of experiment purports to show

contrast, in the case of a cue only the receiver has specifically designed mechanisms. Finally, in the case of coercion it is only the sender that has specifically evolved mechanisms (see, e.g. Maynard Smith & Harper, 2003; Scott-Phillips, 2008). To what category do facial expressions of emotions belong? A priori, all three possibilities are open. However, a good case can be made that they are in fact signals. The argument is easy to make for the senders. While facial expressions of emotions are often related to non-communicative needs (such as having eyes wide open in the presence of a fearful stimuli), they are typically exaggerated or have features that are hard to explain in purely non-communicative terms (blushing, smiling, etc.). What of the receivers? Experiments have revealed that people have two systems to process emotional expressions. The first is designed to react to the expression as cues: a disgust face is treated as a disgusting stimulus. By contrast, the second is sensitive to the communicative nature of the facial expression: the fact that it tells the received that the sender is experiencing disgust (see for instance Ruys & Stapel, 2008). While this might not be considered a conclusive proof, it is plausible that at least some facial expressions or emotions are signals, and we will treat them as such here.

that our mental systems begin by automatically accepting any communicated information before examining it and, potentially, rejecting it (Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993, see also Recanati, 1997, for a theoretical argument). In a representative experiment, participants were told they would have to learn Hopi words. They were then presented with sentences such as “A monishna is a star”, followed shortly by TRUE or FALSE, indicating the veracity of the preceding sentence. In some cases, however, participants were distracted during their processing of the TRUE or FALSE indication. Later, they were subjected to a recognition task: the same statements about Hopi words would be presented, and the participants had to determine whether they were true or false. The authors predicted that if disembedding is not automatic, then interruptions of the TRUE indication would tend to produce more recognition as being false. On the other hand, if disembedding is automatic, then the opposite pattern should emerge: interrupted FALSE indications should lead to remember the statement as being true. Supporting the author’s hypothesis, the latter pattern of results was observed, leading them to conclude that “you can't not believe everything you read” (Gilbert et al., 1993). But, while these results seem to contradict the logic of good design for filtering mechanisms, several caveats are in order.

More recent experiments have highlighted the many shortcomings in the materials used by Gilbert and his colleagues. First, the participants have no previous knowledge that could help them evaluate the sentence. When this is not the case—particularly when the sentences contradict some previous beliefs of the participants—the effect disappears completely and performance is near ceiling (Richter, Schroeder, & Wöhrmann, 2009). Second, the statements used in the original experiments become completely irrelevant if false (knowing that “A monishna is not a star” is quite useless), which is not true of much of the information we encounter. For instance, knowing that “Patrick is a good father” is false is very informative. So while participants who discover that “A monishna is a star” is false have no motivation to remember this information, participants discovering that "Patrick is a good father" is false should be motivated and able to remember that Patrick is not a good father. This is precisely what happens: the original effect all but disappears if the sentences used are informative when false (Hasson, Simmons, & Todorov, 2005). Finally, in the original experiments the source of the information is both trustworthy and quite irrelevant. There is no a priori reason to doubt a computer spouting out translations of Hopi words, and there is very little interest in learning about the truthfulness of this computer. This is exceedingly rare outside the psychology laboratory. Most of the information we get is from people about whom we have a great deal of information (we even track the reputation of journalists, or at least of newspapers or news networks), and whose lies or mistakes are highly relevant (can I trust this friend, this colleague, this newspaper?). Accordingly, when the source is specified and is relevant (the best friend of the participant), then the effect, once again, disappears (Bergstrom & Boyer, submitted).

We can conclude from these experiments that outside the psychology laboratory, disembedding is never—or nearly never—automatic[4]. By contrast, non-ostensive communication need not benefit from the protection offered by the metarepresentational format of ostensive communication. For instance, if I become afraid when I see a frightened individual, the format of the mental representation elicited in the receiver is not of the form "sender intends me to believe that there is something frightening", but of the form "there is something frightening". By getting rid of a layer of protection—the metarepresentational

[4] One proviso should be mentioned: in the experiments conducted so far, the source has never been someone very close to the participant, someone who the participant could trust—nearly—completely, such as a close parent or, in some domains, a teacher. It could therefore be argued that in such cases disembedding is automatic. However, Bergtrom and Boyer (submitted) have found that disembedding is not automatic when the statements come from a close friend. Moreover, the reason why this is the case might be that if a close friend is caught telling something false, this is much more relevant than if it is a stranger that is thus caught. So it would seem reasonable to speculate that such a result would also be obtained with, say, a family member.

embedding—non-ostensive communication could be abused much more easily by senders. Following the logic of the present argument, it should yield less influence than ostensive communication[5]. Non-ostensive communication, however, should still have much more influence than subliminal stimuli because of one major difference: the source is identifiable. Knowing the source provides at least two means to ensure that communicated information can be safely accepted. One is that it is possible to punish a misleading source, provided we find out we have been misled and we can remember who was responsible. Such a punishment can vary widely in form, from direct physical harm to gossip or even simply making it less likely to believe the person in the future. Even though the later is not intuitively construed as punishment, it can make communication harder for the source, which can certainly exert a certain cost. Whatever form it takes, punishment increases the costs for senders of communicating misleading information, making them less likely to do so. But the main advantage offered by the knowledge of a source of information is that we can use our previous knowledge about the source to evaluate whether we should accept what she communicates: for instance, has she been reliable in the past?

These considerations lead to the following predictions: (i) non-ostensive communication should have more influence than subliminally presented stimuli but (ii) less than ostensive communication, and (iii) people should rely both on compatibility with their previous goals and on knowledge about the source to evaluate non-ostensive communication. In what follows, the focus will be on emotional signals (such as facial displays of anger, fear, etc.) as examples of non-ostensive communication.

[5] This generalization can easily accept exceptions. In particular, some emotional contagion carries little costs if one is deceived, but important costs for incredulousness. Panic could be an example: the costs of panicking when one shouldn’t are much lower than those of not panicking when one should.

Point (i) is relatively trivial. Various emotional displays can exert an influence through communication: blushing can make us decide to pursue further a romantic interest, a child seeing his mother's angry face can start preparing an excuse for these missing cookies, etc. Nothing even remotely as effective has been demonstrated with subliminally presented faces: such stimuli may elicit an increase in amygdala activation (Morris, Öhman, & Dolan, 1998; Whalen et al., 1998) and some minor facial movements (Dimberg, Thunberg, & Elmehed, 2000), but not more. This is related to the fact that subliminally presented conditioned stimuli elicit much lower responses than supraliminal conditioned stimuli in general (Olsson & Phelps, 2004).

Point (ii) may be less obviously true, but the examination of a few cases should be sufficient to make the point. You are watching someone on TV. He is smiling and his voice reflects his happy mood. If you have no reason to dislike him, this can certainly be quite enjoyable: moods can be transmitted through such signals (e.g. Neumann & Strack, 2000). Now it's November the fourth, 2008, and the man is Barack Obama. He's delivering his victory speech. For millions of people, this moment is not simply enjoyable, it is fraught with an emotional intensity that will forever be stamped in their mind. Had Obama's smile been meant to convey that he had won the election, this simple gesture, now ostensive, could have provoked infinitely more emotion than that procured through mere contagion. Cultural productions reflect the power of ostensive communication: we read uplifting stories instead of listening to the prosody of someone in a good mood and we watch horror movies instead of looking at scared faces. It would be hard to deny that our most intense emotions have either come from events in our own lives or from stories told by people we hold dear, rather than through watching facial displays or other such non-ostensive signals.

Finally, point (iii) is probably the most contentious. Contrary to all the talk about 'automaticity' in non-ostensive signals such as facial displays of emotions (which is often understood as mandatoriness in the sense of Fodor, 1983), the present theory predicts that the reaction of individuals faced with such signals should be heavily modulated by their source. While previous goals should still play a role in the behavioral outcome, potential conflicts between these goals and the communicated information can now be overcome if the source is deemed to be reliable enough. Unfortunately, while there is a wealth of information about how adults (Petty & Wegener, 1998) and children (Clément, in press; Harris, 2007; Mascaro & Sperber, 2009) use the source of ostensive communication in their evaluation of it, there is much less research on source effects for non-ostensive communication. Still, the little there is points towards a strong and reliable effect of the source. Empathic blushing—due to vicarious embarrassment—is observed when someone is looking at oneself (on tape) or a friend doing something embarrassing; but the effect is much reduced if it is a stranger performing the same action (Shearn, Spellman, Straley, Meirick, & Stryker, 1999). We 'automatically' imitate the facial expression of anger (or other negative emotions) of people belonging to positively valued groups, but not of people from other groups (Bourgeois & Hess, 2007; Mondillon, Niedenthal, Gil, & Droit-Volet, 2007). Responses to facial displays of pleasure or pain are often empathic but they can become *counter*-empathic (smiling in reaction to an expression of pain, and vice versa) if we expect to compete with the other person (Lanzetta & Englis, 1989). If someone has been unfair to us, for some of us (the males) it is the reward, and not the empathy, system that will be activated when that individual is hurt (Singer et al., 2006)[6]. Evaluation of the source can also be more fine-grained than a simple good or bad: political attitudes towards leaders influence reactions to their emotional displays (McHugo, Lanzetta, & Bush, 1991).

[6] Simply seeing someone experiencing an event that should cause pain is not typically part of communication. However, to the extent that empathy could easily be manipulated (by faking pain), it makes sense that individuals should be wary of empathizing with others not deemed to be reliable (and, indeed, mice [(Langford et al., 2006)] and monkeys [(Masserman, Wechkin, & Terris, 1964)] will empathize only with selected individuals). See De Vignemont and Singer, 2006, and Hein and Singer, 2008, on the modulation of empathic responses.

That emotional contagion is modulated by its source is also clearly apparent in the way it contributes to spreading emotions. If emotional contagion were truly automatic, emotions should spread like wildfire (as implicitly argued by, e.g., Christakis & Fowler, 2009). And sometimes it is exactly what seems to be happening in instances of "mass hysteria": whole schools are erupting with inexhaustible laughter (e.g. Ebrahim, 1968) and factories are swept with strange emotional symptoms (e.g. Stahl & Lebedun, 1974). While these observations lend superficial support to the automatic view of emotional contagion, a deeper examination mostly highlight its limits. First of all these episodes, remarkable as they may be, are still relatively rare: their scarcity is as much to be explained as their existence, and it is hard to explain from a purely automatic point of view. Second, echoing the experimental findings reviewed above, the spread of the emotions is heavily modulated by the source of the displays. The emotions and the behaviors they generate do not spread to people who might be brought in to investigate the case, or to other strangers. Instead, they spread among individuals who know and are close to each other: pupils in a classroom, workers in the same plant. Finally, these apparently strange behaviors always seem to serve a goal of the individuals, even if this goal is unconscious. For instance, "mass hysteria" in work settings typically arises when the workers are "suddenly exposed to what they perceive to be an imminent threat" (Evans & Bartholomew, 2009, 364). It is thus not surprising to find a very good correlation between the severity of the symptoms and the level of job dissatisfaction (Stahl & Lebedun, 1974). Likewise, mass hysteria in schools typically targets "a socially cohesive group … exposed to a stressful stimuli" (Evans & Bartholomew, 2009, 384). So, far from being a testimony to the automaticity of emotional contagion, such cases of "mass hysteria" show the transmission of emotions to be modulated by the predicted factors: source and compatibility with previous goals.

Emotion contagion is far from being the only kind of non-ostensive communication in humans. It is a convincing example, however, because it is generally thought to be automatic and therefore powerful. According to the present argument, it is precisely *the opposite* that is the case: the limited influence non-ostensive communication has is made possible by its non-automaticity (more precisely, by the fact that it is modulated by the source). Even less automatic mechanisms, such as ostensive communication, yield even more influence.

Are smart people gullible?

I started this review by peeling of as many layers of filtering mechanisms as possible. When people have to rely on the most primitive of filtering mechanisms, communicated information barely has any influence on them. When people have more cues to the reliability of communicated information—such as its source in the case of non-ostensive communication—their filtering mechanisms allow for more ample influence. This influence is still less, however, than that observed when people have the luxury of examining communicated information while it is embedded in a metarepresentational context before deciding whether they should accept it. This is this last type of communication—ostensive communication—that is by far the most important in humans. It is therefore to be expected that several filtering mechanisms should be dedicated precisely to ostensively communicated information. In this last step in the process, I will look at reasoning as one of the latest addition to the list of filtering mechanisms and try to see what happens when we peel it off. In line with the hypothesis laid out at the outset, if this layer of filtering mechanism is removed, then people should accept less information. But in order to show that, I must first make that case that reasoning is indeed such a filtering mechanism.

The content of communicated information (”Is it compatible with my plans and beliefs?”) as well as its source (”Is she reliable?”) are routinely used in the process of evaluation. Yet these mechanisms still lead to the rejection of some potentially beneficial information. Sometimes we are wrong; our beliefs are mistaken, our plans suboptimal. When this is the case, we would be better off being influenced by others. Trust can help solve this problem by allowing people to override basic compatibility checking when information comes from a reliable source. But trust is far from perfect. For instance, it can be very long in the making, but it is easily lost (Slovic, 1993). In many cases, people would benefit from accepting a piece of information even when trust alone would lead to its rejection. A solution to this dilemma is for the sender to provide reasons supporting the information she wishes to communicate, reasons that can then be examined by the receiver. For instance, Mary might not believe Peter when he tells her that John is a womanizer: she does not know Peter that well, and she previously had a rather positive opinion of John. Without the ability to provide reasons, the situation would stall and the evil John might woo Mary. But if Peter is able to find reasons, he could support his statement with additional information such as: “He slept with Rita and then never called back” (a piece of information Mary can check with her friend) or “Remember the way he flirted with Sarah even though his girlfriend Britney was there?” (a event from which she might have made no inference). Mary can then evaluate these arguments and, if they are deemed sufficient, change her mind about John.

The ability to find good arguments, however, does not come for free (Mercier, in press-b). Even for a skilled language user, the ability to make someone else *understand* our statements is not the same as the ability of making someone *accept* these statements. In the example above, Peter can easily lead Mary to understand that he wants her to believe that John is a womanizer. Given that the skills required to perform this task have done their duty perfectly (Mary understood what Peter meant), they would be of no use in the next step—acceptance. For this, a new cognitive device (or a set of devices) is required. Given that the task of this device will be to find and evaluate reasons, we can call it "reasoning" (Mercier & Sperber, 2009).[7]

This view of reasoning as a specific mental mechanism is very much in line with a host of empirical work in psychology that goes under the umbrella of "dual process theories" (Evans, 2008; Kahneman, 2003). According to these theories, the mind can usefully be divided into two broad categories of processes: intuitions, and reasoning[8]. A theoretically grounded way to frame this distinction is to express it in terms of intuitive and reflective inferences (Mercier & Sperber, 2009). The vast majority (in humans) or the whole (in other animals) of inferences are intuitive: they are performed without attending to the reasons why they are performed. For instance, we very quickly form impressions of the people we meet (e.g. Fiske, Cuddy, & Glick, 2007). These impressions are barely ever based on a conscious, explicit assessment of reasons. Instead they are influenced by factors that we would be hard put to introspect (the shape of the face) or that we would even disown if we knew about their effect (the color of the skin). On the other hand, we are sometimes willing, and able, to revise these initial assessments on the grounds of reasons we consciously ponder. For instance, we might decide that our assessment of an individual was biased by some factor—gender, ethnicity, clothing—and try to revise it accordingly. Faced with visitors from a foreign culture, we often have to remind ourselves that 'strange' behaviors that might intuitively lead us to a grim assessment are to be evaluated in light of different norms.

[7] This assumes that understanding and acceptance are two distinct processes, and acceptance is not taken for granted for understanding. This runs against some views of language comprehension that rely on 'interpretive charity'—i.e. people need to assume most statements to be true to be able to understand them (Davidson, 1984). A forceful critique of these views from a perspective congenial to the one adopter here can be found in (Sperber et al., in press).

[8] While most authors use reasoning to describe the two levels, describing them as "system 1 and system 2" (Stanovich, 2004), "heuristic and analytic" (Evans, 2007), or "associative and rule-based" (Sloman, 1996) here we follow Kahneman, 2003, and Mercier and Sperber, 2009, in using intuition for the first type of mechanisms and restricting reasoning to the second.

Following the Cartesian tradition, standard dual process theories tend to ascribe to reasoning a mostly individual function: by allowing us to correct for the errors of our intuitions, reasoning should lead us towards better decisions and epistemic improvement (Evans & Over, 1996; Kahneman, 2003; Stanovich, 2004). This is where the view presented here parts way with classical models. Instead of seeing reasoning as a prop of individual cognition, the argumentative theory suggests that reasoning evolved for a profoundly social purpose: finding reasons to convince other people and evaluate these reasons so as to be convinced only when we should be (Mercier & Sperber, in press).[9] In this perspective, reasoning is one of the most recent—if not the most recent—layer of filtering mechanism to have been added to our cognitive makeup. In line with the argument offered at the beginning, it should therefore allow people to be *more* influenced by communication. This prediction clashes with the Cartesian view of the good reasoner as the über-skeptic, rejecting naïve beliefs through constant doubt and careful examination. In what follows I will offer a defense of the present view in four points: (i) Despite the confirmation bias, people can be swayed by good arguments, even when their conclusions conflict with some previous beliefs; (ii) reasoning does a better job than other mechanisms at transmitting counterintuitive beliefs; (iii) trying to stop people from reasoning does not lead to good results in terms of influence; (iv) reasoning more leads to more diverse beliefs, often including more false beliefs.

(i) According to the argumentative theory, reasoning is designed in part to produce arguments to convince other people. Reasoning should find reasons that support one's point of view or rebut the interlocutor's, and not the opposite. A confirmation bias is therefore to be expected. And, indeed, people exhibit a strong and robust confirmation bias (Nickerson, 1998). While such a bias is predicted by the hypothesis, it could be problematic for the

[9] There is no space here to present, even briefly, the empirical evidence that supports the theory, but it is comprehensively presented in the reference mentioned. Other evidence can be found in Mercier (in press-a), which defends the idea that argumentation is a human universal, and Mercier (submitted), which shows that children are skilled arguers from very early on.

present argument. The prevalence of the confirmation bias could lead to the conclusion that people will be bad at evaluating arguments. Indeed, many experiments have shown that people are biased in their evaluation of arguments, being more prone to discover flaws or find counterarguments when the conclusion disagrees with their own opinion (see (Mercier & Sperber, in press). In some cases, their prior beliefs completely skew the way they treat arguments, to the point that being presented with contradictory evidence can even strengthen these prior beliefs (Batson, 1975; Burris, Harmon-Jones, & Tarpley, 1997; Lord, Ross, & Lepper, 1979; Tormala & Petty, 2002). One could think that, because of the confirmation bias, reasoning leads people to be *less* influenced by communication, not more—they become more pigheaded once again, contrary to the earlier prediction. It is therefore important to emphasize that this is, in fact, not the case.

Two factors prevent us from drawing strong negative conclusions from the studies mentioned above. First, in most cases, despite a biased assessment, the arguments—to the extent that they are strong—still change the participants' attitudes (Petty & Wegener, 1998). Experiments reporting polarization of prior beliefs following counterarguments depend on a very specific set of circumstances, rarely obtained (Kuhn & Lao, 1996; Miller, Michoskey, Bane, & Dowd, 1993; Pomerantz, Chaiken, & Tordesillas, 1995). Second, these experiments never put participants in the context of a dialogue. Instead, the participants are faced with a (generally) written argument and its evaluation is soon succeeded by a phase in which the participants produce counterarguments (when they do not agree with the conclusion). But there is no one to refute these counterarguments or to propose new arguments for the other side. In such an artificial context, the confirmation bias present in the production of arguments can proceed unimpeded and make the evaluation appear much more biased than it actually is. If the participants were dealing with an interlocutor able to defend the opposing point of view, they would not have such leisure to be biased. Accordingly, the confirmation bias can be much attenuated in group settings (Kuhn, Shaw, & Felton, 1997). Moreover, in many cases members who hold a minority view in a debate are able to prevail—for instance when they have understood a reasoning problem (Laughlin & Ellis, 1986; Moshman & Geil, 1998). For this to be the case, all the other group members must have been convinced to change their views, in spite of any confirmation bias they might have: even though they certainly try, at first, to defend their views, they run short of argument at some point and come to accept the correct answer.[10] We can conclude that the confirmation bias does not stop people from changing their minds when they are confronted with good arguments.

(ii) In a more historical perspective, it is possible to take the existence of science as a testimony to the power of reasoning to change people's mind, even to the point where they come to accept counterintuitive beliefs. If one might claim that most of us acquire our scientific knowledge (outside our area of expertise for scientists) through trust in authority, such was clearly not the case when this knowledge started to emerge. The history of science is a long uphill battle in which people become convinced of otherwise weird things by good arguments and evidence. Indeed, most of the scientific knowledge we have now is deeply counterintuitive: continents move, time changes with speed, life emerged out of inert chemicals, the universe has more than three dimensions, etc. (see Cromer, 1993). We should therefore expect people to exhibit a strong confirmation bias in such cases. While this has no doubt been the case—scientists are far from being exempt from this bias (Mahoney, 1977; Nickerson, 1998)—it has not stopped them from, eventually, accepting the arguments and the conclusions they support, counterintuitive as these conclusions may be. Other groups that

[10] Lately, the capacity of deliberation to revise people's attitudes has also come under attack in political science. According to some critics of deliberative democracy, debates among groups of citizens are rather futile and only rarely succeed in changing minds (partly because of the confirmation bias) (Goodin & Niemeyer, 2003; Hibbing & Theiss-Morse, 2002; Sunstein, 2002). This conclusion, however, mostly stems from a misinterpretation of the empirical data (Mercier & Landemore, submitted) and problems in the way the effects of debates are measured (Mackie, 2006).

mostly rely on reasoning to transmit their beliefs have arrived at very counterintuitive constructs. Mathematicians have created objects that we only a few people can represent, logicians have made conditionals yield paradoxical conclusions, philosophers have developed postmodern relativism and economists can sustain a strange faith in the free market. By comparison, beliefs that spread through other means tend to be much more intuitive. It can be argued that religious beliefs, for instance, spread mostly through trust, whether it is in one's parents, one's friends, or in a more official representative of a given religion (see for instance (Stark & Bainbridge, 1980). Accordingly, religious beliefs can be counterintuitive, but not too much (Boyer, 1996a, 1996b). Even within religion, it might be that the most counterintuitive beliefs are the ones arrived at through painstaking theological argument, and that these beliefs require more reasoning to spread. The historical evidence could thus be interpreted as showing that reasoning is more effective in making people accept deeply counterintuitive beliefs than other means of communication.

(iii) Given the widespread idea that reasoning is generally used to resist influence, it should come as no surprise that people have tried to deprive individuals of their ability to reason in order to get them to accept some beliefs. Unfortunately, some of the people sharing this view had total control over other people and no scruple. During the war in Korea, 7,190 Americans were held captive by the Korean and the Chinese. Many of them were subject to intense 'brainwashing', from long daily (or twice-daily) lectures on the benefits of communism and the pitfalls of capitalism to group discussions on the same topic (Jowett, 2006). Possibly in order to make the prisoners more receptive to the arguments, they were treated very harshly, deprived of sleep and underfed: not the best conditions for reasoning. Yet only 21 soldiers defected and chose to go to China after the war—not exactly a success[11].

[11] In fact, the term "brainwashing" started to be used when the defection of the US POWs became an embarrassment to their officers. These officers were very happy to have found an excuse for this apparently treacherous behavior. The term had been invented for other

Moreover, these prisoners later recanted, came back to the US and said that they had done this to put an end to their terrible treatment (Streatfeild, 2007). This null rate of success is to be compared with the two following numbers. First, 2,730 POWs died while in captivity Killing people is apparently easier than changing their minds. This shows how hard it is to convince people without trust or good arguments. Second, 50,000 Chinese and Korean POWs decided to defect without the help of any brainwashing. This comparison demonstrates that an actual display (the US offers a much superior material comfort) can be much more effective than communication.

People have also charged some new religious movements with brainwashing their recruits, and using techniques to deprive them of their reasoning ability, from sleep deprivation to overworking. Yet these new religious movements encounter, on the whole, very little success: most of their recruits choose to leave the cults/religions quite quickly and their growth rate has been close to zero (Anthony & Robbins, 2004; Streatfeild, 2007). By comparison, sects such as the early Christians or the Mormons have been much more successful using normal means of influence: gaining people's trust little by little, reaching people through their friends and families, etc. (Stark, 1996; Stark & Bainbridge, 1980). Techniques associated with brainwashing can also be used by interrogators to obtain compliance from suspects. Yet, here again, when the suspect is highly motivated and intelligent, the only chance is to talk to him and get to a position where a modicum of trust has been established and well targeted arguments can be used (Streatfeild, 2007, 375; Alexander & Bruning, 2008). On the whole, trying to reduce people's capacity to reason has proven to be a very inefficient—indeed, often counterproductive—strategy.

propaganda purposes a few years earlier by a CIA propagandist working undercover as a journalist, Edward Hunter.

Besides the evolutionary argument, there also is a good mechanistic reason why stopping people from reasoning will rarely make it easier to convince them. When people use reasoning to evaluate arguments, part of the process is a search for counterarguments. Conviction will then be stronger if the attempt is unsuccessful. But if someone is prevented from reasoning, she will not be able to engage in such a search for counterarguments. As a result, she is much less likely to be convinced by the argument. Now, given that arguments are typically offered in support of conclusions that would not otherwise be accepted, this means that the receiver will revert to her earlier evaluation and reject the conclusion. This is a very good design for a filtering mechanism: if it is disrupted, it reverts to a negative assessment; only its good functioning can allow more information to be accepted. Experimental results confirm this prediction: people who think more about good arguments tend to be more convinced by them, probably because they did not manage to come up with good enough counterarguments (if they had, they would not change their mind, however good the original argument might be) (e.g. Cacioppo, Petty, & Morris, 1983; Petty & Cacioppo, 1979), and, of particular interest to academics, this study showing that people who have had a coffee are more alert and change their mind more in response to good arguments (Martin, Laing, Martin, & Mitchell, 2005).

(iv) Both the classical view of reasoning and the argumentative theory agree that on the whole more or better reasoning should be conducive to more accurate beliefs. Where the two theories differ is in the way this result is achieved. In the classical view, one of the most important uses of reasoning is to critically evaluate our own beliefs as well as those of others. In this case, epistemic improvement is mostly achieved through a weeding out of unjustified beliefs, resulting in a higher share of justified beliefs. According to the argumentative theory, on the other hand, reasoning achieves epistemic improvement by allowing us to accept more justified beliefs, mostly stemming from communication. As a result, even if the predicted share of true beliefs were the same in both cases, the quantity of beliefs that the use of reasoning leads to would be different: fewer beliefs for the classical view, more according to the argumentative theory.[12] In turn, this difference in quantity must logically give rise to a difference in diversity. Thus, according to the argumentative theory, people who reason more, or better, should have more diverse beliefs. It might then be expected that people who reason more have *more* false beliefs than other people, but that they also have *even more* true beliefs than them.

A cursory examination of the beliefs of people with impeccable reasoning credentials seems to confirm this prediction. From Newton, who spent more time on biblical numerology than on physics, to Linus Pauling who was convinced that vitamin C was the ultimate secret to a healthy life, brilliant intellectuals have had the habit of accepting weird beliefs. Some paranormal beliefs correlate positively with variables usually associated with more or better reasoning (such as cognitive ability or level of education) (Tobacyk, Miller, & Jones, 1984). Mensa members seem to be particularly prone to belief in extra-sensory perception (Shermer, 1997). Smart, educated people are often among the early adopters of novel beliefs (see Vyse, 1997, in the case of New Age, Shermer, 1997, in the case of UFOs and alien abductions, and Wallace, 2009, in the case of rejection of vaccination). While such evidence falls short of being conclusive, it makes it hard to deny that good reasoning is very far from being foolproof and that even the best reasoners have a tendency to accept weird beliefs that quite often turn out to be wrong.

[12] The strength of this conclusion rests on how the classical view construes reasoning. If it construes it as working mostly through self-criticism, then the difference between the predictions of the two views should be large. On the other hand, if reasoning, in the classical view, also relies significantly on the construction of new beliefs, then the gap between the predictions of the two views is smaller. Still, as long as there is a difference in the use of our critical abilities between the two views—if they are aimed mostly at ourselves for the classical view, and mostly at communicated information for the argumentative view—then this conclusion should hold.

The four points above converge towards the conclusion that the more people reason, the more they can be influenced: Despite the confirmation bias, people can change their mind when faced by good arguments, even to the point of accepting counterintuitive beliefs; People deprived of reasoning are very hard to influence while people who rely a lot on reasoning often entertain a wider range of beliefs, including quite a few wrong ones.

Conclusion

This chapter set out to explore some of the filtering mechanisms we use when dealing with communicated information. It suggested that these mechanisms would have evolved in successive layers, allowing communicated information to be increasingly influential. The following claims were defended:

- When we have very limited means of evaluation, as in the case of subliminal influence, we revert to old mechanisms of competition between goals that leave communicated information with only small effects, and only when those effects are compatible with our previous goals.
- Because non-ostensive communication can, and does, use the source of information to modulate its effect, it can yield more influence than subliminal influence. However, it does not enjoy the layer of protection offered by ostensive communication, and so it is much less influential.
- Reasoning is a recently evolved filtering mechanism. By allowing us to understand and evaluate arguments, reasoning makes it possible to communicate beliefs that would otherwise have very little chance of being accepted by receivers: it increases the amount of information efficiently transmitted.

From an evolutionary perspective, the results presented here should come as no surprise. We know that communication exerts vastly more influence in humans than in any other primate species. It is reasonable to assume that this influence has been ever growing during our evolution. Given that the effects of communication must be held in check by filtering mechanisms, it follows that these mechanisms should have evolved to allow for this increasing influence of communicated information.

While they may not be surprising from such a point of view, these results still run against some widely shared intuitions: that less cognitively gifted people should be easier to influence, that preventing people from reasoning should make them more manipulable, etc. These intuitions explain the success—in popular opinion, not in fact—of subliminal influence. Subliminal influence may have been harmless, but other manipulation attempts based on these intuitions took a much more somber aspect, from sleep deprivation to sensory isolation. It is therefore quite comforting to think that such attempts have little chance of encountering regular success (as the CIA has found out after a great many experiments, see Streatfeild, 2007). On the contrary, the most efficient way to influence people is to win their trust and to use cogent arguments.

Acknowledgments

I wish to thank Brett Calcott, Ben Fraser, Richard Joyce, Dan Sperber and Kim Sterelny for insightful comments on this chapter.

References

Alexander, M., & Bruning, J. 2008. *How to Break a Terrorist: The U.S. Interrogators Who Used Brains, Not Brutality, to Take Down the Deadliest Man in Iraq*. New York: Free Press.

Anthony, D., & Robbins, T. 2004. Conversion and 'brainwashing' in new religious movements. In *The Oxford Handbook of New Religious Movements,* ed. J. R. Lewis, 243-297. Oxford: Oxford University Press.

Bargh, J. A. 2002. "Losing consciousness: Automatic influences on consumer judgment, behavior, and motivation." *Journal of Consumer Research,* 29(2), 280-285.

Batson, C. D. 1975. "Rational processing or rationalization?: The effect of discontinuing information on stated religious belief." *Journal of Personality and Social Psychology,* 32(1), 176–184.

Beall, S. P., Langley, D. J., & Edwards, D. H. 1990. "Inhibition of escape tailflip in crayfish during backward walking and the defense posture." *The Journal of Experimental Biology,* 152, 577.

Bergstrom, B., & Boyer, P. Submitted. "Who mental systems believe: Effects of source on judgments of truth."

Berridge, K. C., & Winkiehnan, P. 2003. "What is an unconscious emotion?(The case for unconscious" liking")." *Cognition and Emotion,* 17 (2), 181-211.

Bourgeois, P., & Hess, U. 2007. "The impact of social context on mimicry." *Biological Psychology*.

Boyer, P. 1996a. Cognitive limits to conceptual relativity: the limitingcase of religious categories. In *Rethinking Linguistic Relativity,* ed. J. Gumperz & S. Levinson, 203–231. Cambridge, MA: Cambridge University Press.

Boyer, P. 1996b. "What makes anthropomorphism natural: Intuitive ontology and cultural representations." *Journal of the Royal Anthropological Institute,* 2(1), 83–97.

Burris, C. T., Harmon-Jones, E., & Tarpley, W. R. 1997. "By faith alone: Religious agitation and cognitive dissonance." *Basic and Applied Social Psychology,* 19(1), 17–31.

Cacioppo, J. T., Petty, R. E., & Morris, K. J. 1983. "Effects of need for cognition on message evaluation, recall, and persuasion." *Journal of personality and social psychology,* 45(4), 805–818.

Carlson, J. M., & Doyle, J. 2002. "Complexity and robustness." *PNAS,* 19(99), 2538-2545.

Christakis, A. N., & Fowler, J. H. 2009. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives* New York: Little, Brown and Company.

Clément, F. in press. "To trust or not to trust? Children’s social epistemology." *Review of Philosophy and Psychology*.

Cromer, A. 1993. *Uncommon Science: The Heretical Nature of Science*. New York: Oxford University Press.

Davidson, D. 1984. Radical interpretation. In *Inquiries into Truth and Interpretation.,* ed. D. Davidson, 125-140. Oxford: Clarendon Press.

Dawkins, R., & Krebs, J. R. 1978. Animal signals: Information or manipulation? In *Behavioural Ecology: An Evolutionary Approach,* ed. J. R. Krebs & N. B. Davies,282-309. Oxford: Basil Blackwell Scientific Publications.

De Vignemont, F., & Singer, T. 2006. "The empathic brain: how, when and why?" *Trends in Cognitive Sciences,* 10(10), 435–441.

Dijksterhuis, A., & Bargh, J. A. 2001. The perception-behavior expressway. In *Advances in experimental social psychology,* ed. M. P. Zanna, Vol. 33, 1-40. San Diego, CA: Academic Press.

Dimberg, U., Thunberg, M., & Elmehed, K. 2000. "Unconscious facial reactions to emotional facial expressions." *Psychological Science*, 86–89.

Ebrahim, G. J. 1968. "Mass hysteria in school children. Notes on three outbreaks in East Africa." *Clinical pediatrics,* 7(7), 437.

Edwards, D. H., Heitler, W. J., & Krasne, F. B. 1999. "Fifty years of a command neuron: the neurobiology of escape behavior in the crayfish." *Trends in Neurosciences,* 22(4), 153-161.

Evans, H., & Bartholomew, R. 2009. *Outbreak! The Encyclopedia of Extraordinary Social Behavior*. New York: Anomalist Books.

Evans, J. S. B. T. 2007. *Hypothetical Thinking: Dual Processes in Reasoning and Judgment*. Hove: Psychology Press.

Evans, J. S. B. T. 2008. "Dual-processing accounts of reasoning, judgment and social cognition." *Annual Review of Psychology,* 59, 255-278.

Evans, J. S. B. T., & Over, D. E. 1996. *Rationality and Reasoning*. Hove: Psychology Press.

Farrell, J., & Rabin, M. 1996. "Cheap talk." *Journal of Economic Perspectives,* 10, 110–118.

Fiske, S. T. 1993. "Controlling other people: The impact of power on stereotyping." *American Psychologist,* 48, 621–621.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. 2007. "Universal dimensions of social cognition: warmth and competence." *Trends in Cognitive Sciences,* 11(2), 77-83.

Fodor, J. 1983. *The Modularity of Mind*. Cambridge, Massachusetts: MIT Press.

Gilbert, D. T., Krull, D. S., & Malone, P. S. 1990. "Unbelieving the unbelievable: Some problems in the rejection of false information." *Journal of Personality and Social Psychology,* 59(4), 601-613.

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. 1993. "You can't not believe everything you read." *Journal of Personality and Social Psychology,* 65(2), 221-233.

Goodin, R. E., & Niemeyer, S. J. 2003. "When does deliberation begin? Internal reflection versus public discussion in deliberative democracy." *Political Studies,* 51(4), 627–649.

Greenwald, A. G., Spangenberg, E. R., Pratkanis, A. R., & Eskenazi, J. 1991. "Double-Blind Tests of Subliminal Self-Help Audiotapes." *Psychological Science,* 2(2), 119-122.

Haig, D. 1993. "Genetic conflicts in human pregnancy." *Quarterly Review of Biology*, 495–532.

Harris, P. L. 2007. "Trust." *Developmental Science,* 10, 135-138.

Hasson, U., Simmons, J. P., & Todorov, A. 2005. "Believe it or not: On the possibility of suspending belief." *Psychological Science,* 16(7), 566-571.

Hein, G., & Singer, T. 2008. "I feel how you feel but not always: the empathic brain and its modulation." *Current Opinion in Neurobiology,* 18(2), 153–158.

Hibbing, J. R., & Theiss-Morse, E. 2002. *Stealth Democracy: Americans' Beliefs about How Government Should Work*. Cambridge, UK: Cambridge University Press.

Jackendoff, R. 1999. "Possible stages in the evolution of the language capacity." *Trends in Cognitive Science,* 3(7), 272-279.

Jowett, G. S. 2006. Brainwashing: The Korean POW controversy and the origins of a myth. In *Readings in Propaganda and Persuasion: New and Classic Essays,* ed. G. S. Jowett & V. O'Donnell, 201-210. Thousand Oaks: Sage Publication.

Kahneman, D. 2003. "A perspective on judgment and choice: Mapping bounded rationality." *American Psychologist,* 58(9), 697-720.

Kitano, H. 2004. "Biological robustness." *Nature Review Genetics,* 5(11), 826-837.

Krasne, F. B., & Lee, S. C. 1988. "Response-dedicated trigger neurons as control points for behavioral actions: selective inhibition of lateral giant command neurons during feeding in crayfish." *Journal of Neuroscience,* 8(10), 3703.

Krasne, F. B., Shamsian, A., & Kulkarni, R. 1997. "Altered excitability of the crayfish lateral giant escape reflex during agonistic encounters." *Journal of Neuroscience,* 17(2), 709.

Krebs, J. R., & Dawkins, R. 1984. Animal signals: Mind-reading and manipulation? In *Behavioural Ecology: An Evolutionary Approach,* ed. J. R. Krebs & N. B. Davies, 390-402. Oxford: Basil Blackwell Scientific Publications.

Kuhn, D., & Lao, J. 1996. "Effects of Evidence on Attitudes: Is Polarization the Norm?" *Psychological Science,* 7, 115-120.

Kuhn, D., Shaw, V. F., & Felton, M. 1997. "Effects of dyadic interaction on argumentative reasoning." *Cognition and Instruction,* 15, 287-315.

Langford, D. J., Crager, S. E., Shehzad, Z., Smith, S. B., Sotocinal, S. G., Levenstadt, J. S., et al. 2006. "Social modulation of pain as evidence for empathy in mice." *Science,* 312(5782), 1967.

Lanzetta, J. T., & Englis, B. G. 1989. "Expectations of cooperation and competition and their effects on observers' vicarious emotional responses." *Journal of Personality and Social Psychology,* 56(4), 543–554.

Laughlin, P. R., & Ellis, A. L. 1986. "Demonstrability and social combination processes on mathematical intellective tasks." *Journal of Experimental Social Psychology,* 22, 177–189.

Lord, C. G., Ross, L., & Lepper, M. R. 1979. "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." *Journal of Personality and Social Psychology,* 37(11), 2098-2109.

Mackie, G. 2006. "Does democratic deliberation change minds?" *politics, philosophy & economics,* 5(3), 279.

Mahoney, M. J. 1977. "Publication prejudices: An experimental study of confirmatory bias in the peer review system." *Cognitive Therapy and Research,* 1(2), 161-175.

Martin, P. Y., Laing, J., Martin, R., & Mitchell, M. 2005. "Caffeine, cognition, and persuasion: Evidence for caffeine increasing the systematic processing of persuasive messages." *Journal of Applied Social Psychology,* 35(1), 160–161.

Mascaro, O., & Sperber, D. 2009. "The moral, epistemic, and mindreading components of children's vigilance towards deception." *Cognition,* 112, 367–380.

Masserman, J. H., Wechkin, S., & Terris, W. 1964. ""Altruistic" behavior in rhesus monkeys." *American Journal of Psychiatry,* 121(6), 584–585.

Maynard-Smith, J. 1994. "Must reliable signals always be costly?" *Animal Behaviour,* 47, 1115-1120.

Maynard Smith, J., & Harper, D. 2003. *Animal Signals*. Oxford: Oxford University Press.

McHugo, G. J., Lanzetta, J. T., & Bush, L. K. 1991. "The effect of attitudes on emotional reactions to expressive displays of political leaders." *Journal of Nonverbal Behavior,* 15(1), 19–41.

Mercier, H. in press-a. "On the universality of argumentative reasoning." *Journal of Cognition and Culture*.

Mercier, H. in press-b. "The social origins of folk epistemology." *Review of Philosophy and Psychology*.

Mercier, H. submitted. "Developmental evidence for the argumentative theory of reasoning."

Mercier, H., & Landemore, H. submitted. "Reasoning is for arguing: Understanding the successes and failures of deliberation."

Mercier, H., & Sperber, D. 2009. Intuitive and reflective inferences. In *In Two Minds,* ed. J. S. B. T. Evans & K. Frankish. New York: Oxford University Press.

Mercier, H., & Sperber, D. in press. "Why do humans reason? Arguments for an argumentative theory." *Behavioral and Brain Sciences*.

Milgram, S. 1974. *Obedience to Authority: An Experimental View*. New York: Harper & Row.

Miller, A. G., Michoskey, J. W., Bane, C. M., & Dowd, T. G. 1993. "The attitude polarization phenomenon: role of response measure, attitude extremity, and behavioral consequences of reported attitude change." *Journal of Personality and Social Psychology,* 64(4), 561-574.

Mondillon, L., Niedenthal, P. M., Gil, S., & Droit-Volet, S. 2007. "Imitation of in-group versus out-group members' facial expressions of anger: a test with a time perception task." *Social neuroscience,* 2(3-4), 223.

Moore, T. E. 1982. "Subliminal advertising: What you see is what you get." *Journal of Marketing,* 46(2), 38-47.

Morris, J. S., Öhman, A., & Dolan, R. J. 1998. "Conscious and unconscious emotional learning in the human amygdala." *Nature,* 393(6684), 467–470.

Moshman, D., & Geil, M. 1998. "Collaborative reasoning: Evidence for collective rationality." *Thinking and Reasoning,* 4(3), 231-248.

Neumann, R., & Strack, F. 2000. "mood contagion: The automatic transfer of mood between persons." *Journal of personality and social psychology,* 79(2), 211–223.

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

Olsson, A., & Phelps, E. A. 2004. "Learned fear of" unseen" faces after Pavlovian, observational, and instructed fear." *Psychological Science,* 15(12), 822.

Petty, R. E., & Cacioppo, J. T. 1979. "Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses." *Journal of Personality and Social Psychology,* 37, 349-360.

Petty, R. E., & Wegener, D. T. 1998. Attitude change: Multiple roles for persuasion variables. In *The Handbook of Social Psychology,* ed. D. Gilbert, S. Fiske & G. Lindzey, Vol. 1, 323–390. Boston: McGraw-Hill.

Pinker, S., & Bloom, P. 1990. "Natural language and natural selection." *Behavioral and Brain Sciences,* 13(4), 707-784.

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. 1995. "Attitude strength and resistance processes." *Journal of Personality and Social Psychology,* 69(3), 408-419.

Pratkanis, A. R., & Aronson, E. 1992. *Age of Propaganda: The Everyday Use and Abuse of Persuasion*. New York: W.H. Freeman and Company.

Recanati, F. 1997. "Can we believe what we do not understand?" *Mind and Language,* 12(1), 84–100.

Richter, T., Schroeder, S., & Wöhrmann, B. 2009. "You don't have to believe everything you read: background knowledge permits fast and efficient validation of information." *Journal of Personality and Social Psychology,* 96, 538–558.

Ruys, K. I., & Stapel, D. A. 2008. "Emotion elicitor or emotion messenger? Subliminal priming reveals two faces of facial expressions." *Psychological Science,* 19(6), 593-600.

Scott-Phillips, T. C. 2008. "Defining biological communication." *Journal of evolutionary biology,* 21(2), 387–395.

Shearn, D., Spellman, L., Straley, B., Meirick, J., & Stryker, K. 1999. "Empathic blushing in friends and strangers." *Motivation and Emotion,* 23(4), 307–316.

Shermer, M. 1997. *Why People Believe Weird Things*. New York: Henry Holt and Company.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. 2006. "Empathic neural responses are modulated by the perceived fairness of others." *Nature,* 439(7075), 466.

Sloman, S. A. 1996. "The empirical case for two systems of reasoning." *Psychological Bulletin,* 119(1), 3-22.

Slovic, P. 1993. "Perceived risk, trust, and democracy." *Risk Analysis,* 13(6), 675-682.

Sperber, D. 1997. "Intuitive and reflective beliefs." *Mind and Language,* 12(1), 67-83.

Sperber, D. 2005. Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In *The Innate Mind: Structure and Contents,* ed. P. Carruthers, S. Laurence & S. Stich.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. in press. "Epistemic vigilance."

Sperber, D., & Wilson, D. 1995. *Relevance: Communication and Cognition*. Oxford: Blackwell.

Stahl, S. M., & Lebedun, M. 1974. "Mystery gas: an analysis of mass hysteria." *Journal of Health and Social Behavior,* 15(1), 44–50.

Stanovich, K. E. 2004. *The Robot's Rebellion*. Chicago: Chicago University Press.

Stark, R. 1996. *The Rise of Christianity: A Sociologist Reconsiders History*. Princeton: Princeton University Press.

Stark, R., & Bainbridge, W. S. 1980. "Networks of faith: Interpersonal bonds and recruitment to cults and sects." *The American Journal of Sociology,* 85(6), 1376-1395.

Sterelny, K. in press. *The Fate of the Third Chimpanzee*. Cambridge, MA: MIT Press.

Strahan, E. J., Spencer, S. J., & Zanna, M. P. 2002. "Subliminal priming and persuasion: Striking while the iron is hot." *Journal of Experimental Social Psychology,* 38(6), 556-568.

Streatfeild, D. 2007. *Brainwash: The Secret History of Mind Control*. New York: Thomas Dunne Books.

Sunstein, C. R. 2002. "The law of group polarization." *Journal of Political Philosophy,* 10(2), 175-195.

Tobacyk, J., Miller, M. J., & Jones, G. 1984. "Paranormal beliefs of high school students." *Psychological Reports,* 55, 255-261.

Tormala, Z. L., & Petty, R. E. 2002. "What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty." *Journal of Personality and Social Psychology,* 83(6), 1298–1313.

Vyse, S. A. 1997. *Believing in Magic: The Psychology of Superstition* New York: Oxford University Press.

Wagner, A. 2005. "Robustness, evolvability, and neutrality." *FEBS Letters,* 579(8), 1772-1778.

Wallace, A. (2009, 19 October). An Epidemic of Fear: How Panicked Parents Skipping Shots Endangers Us All. *Wired*.

Weir, W. 1984. "Another look at subliminal "facts"." *Advertising Age, October,* 15, 46.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. 1998. "Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge." *Journal of Neuroscience,* 18(1), 411.

Zahavi, A., & Zahavi, A. 1997. *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.