(To appear in *Mind and Language,* 2010)

Epistemic Vigilance

DAN SPERBER, FABRICE CLÉMENT, CHRISTOPHE HEINTZ, OLIVIER MASCARO, HUGO MERCIER, GLORIA ORIGGI AND DEIRDRE WILSON

Abstract: Humans depend massively on communication with others, but this leaves them open to the risk of being accidentally or intentionally misinformed. We claim that humans have a suite of cognitive mechanisms for epistemic vigilance to ensure that communication remains advantageous despite this risk. Here we outline this claim and consider some of the ways in which epistemic vigilance works in mental and social life by surveying issues, research and theories in different domains of philosophy, linguistics, cognitive psychology and the social sciences.

Acknowledgements: We are grateful to two anonymous referees for useful suggestions and comments on an earlier version of this article and to the Centre for the Study of Mind in Nature at the University of Oslo for supporting our work.

E-mail Address for correspondence: dan.sperber@gmail.com

1. Introduction

We claim that humans have a suite of cognitive mechanisms for epistemic vigilance, targeted at the risk of being misinformed by others. Here we present this claim and consider some of the ways in which epistemic vigilance works in mental and social life. Our aim is to integrate into a coherent topic for further research a wide range of assumptions developed elsewhere by ourselves or others, rather than to present detailed arguments for each.

Humans are exceptional among animals for both the richness and strength of their cognitive abilities and the extent to which they rely on a wide variety of information communicated by others. These two traits are linked. On the one hand, it would not be possible to rely so heavily on rich communication in the absence of species-specific cognitive abilities, in particular language and advanced mindreading. On the other hand, these individual abilities would not develop or function properly in the absence of cognitive skills, conceptual tools, and background knowledge acquired from others.

How reliable are others as sources of information? In general, they are mistaken no more often than we are – after all, 'we' and 'they' refer to the same people – and they know things that we don't know. So it should be advantageous to rely even blindly on the competence of others. Would it be more advantageous to modulate our trust by exercising some degree of vigilance towards the competence of others? That would depend on the cost and reliability of such vigilance. But in any case, the major problem posed by communicated information has to do not with the competence of others, but with their interests and their honesty. While the interests of others often overlap with our own, they rarely coincide with ours exactly. In a variety of situations,

their interests are best served by misleading or deceiving us. It is because of the risk of deception that epistemic vigilance may be not merely advantageous but indispensable if communication itself is to remain advantageous.

Most human communication is carried out intentionally and overtly: The communicator performs an action by which she not only conveys some information but also conveys that she is doing so intentionally (Grice, 1975; Sperber and Wilson, 1995).[1] For communication of this type to succeed, both communicator and addressee must cooperate by investing some effort: in the communicator's case, the effort required to perform a communicative action, and in the addressee's case, the effort required to attend to it and interpret it. Neither is likely to invest this effort without expecting some benefit in return. For the addressee, the normally expected benefit is to acquire some true and relevant information. For the communicator, it is to produce some intended effect in the addressee. To fulfil the addressee's expectations, the communicator should do her best to communicate true information. To fulfil her own expectations, by contrast, she should choose to communicate the information most likely to produce the intended effect in the addressee, regardless of whether it is true or false.[2]

---

[1] Such human communication is very different from the many forms of animal communication discussed by Dawkins and Krebs (1978) and Krebs and Dawkins (1984). However, some similar evolutionary considerations about costs and benefits to communicators and receivers are relevant to both cases.

[2] Note, however, that if equally desirable effects (including long-term effects on reputation) could be achieved by conveying either true or false information, it might be preferable to communicate true information, since this is generally easier to

There are situations where communicators would not stand to benefit from misleading their audience, for instance when teaching their own children or coordinating joint action. However, humans communicate in a much wider variety of situations, with interlocutors whose interests quite often diverge from their own.

People stand to gain immensely from communication with others, but this leaves them open to the risk of being accidentally or intentionally misinformed, which may reduce, cancel, or even reverse these gains. The fact that communication is so pervasive despite this risk suggests that people are able to calibrate their trust well enough to make it advantageous on average to both communicator and audience (Sperber, 2001; Bergstrom *et al.*, 2006). For this to happen, the abilities for overt intentional communication and epistemic vigilance must have evolved together, and must also develop together and be put to use together. A disposition to be vigilant is likely to have evolved biologically alongside the ability to communicate in the way that humans do. Human social life (with some cultural variability) provides plenty of inputs relevant to the development of psychological mechanisms for epistemic vigilance. Moreover, interaction among epistemically vigilant agents is likely to generate not only psychological but also social vigilance mechanisms. Before examining a variety of these mechanisms, we consider some philosophical issues relevant to the study of epistemic vigilance.

---

manage (see Paglieri and Woods In press, who develop this parsimony argument and draw stronger conclusions from it than we do).

2. Epistemic Trust and Vigilance

Trust is obviously an essential aspect of human interaction (and also an old philosophical topic – see Origgi, 2004, 2008a). What is less obvious is the claim that humans not only end up trusting one another much of the time, but are also trustful and willing to believe one another to start with, and withdraw this basic trust only in circumstances where they have special reasons to be mistrustful. Still, philosophers from Thomas Reid to Tyler Burge and Ruth Millikan, and more recently psychologists such as Daniel Gilbert, have made this stronger claim, arguing that humans are fundamentally trustful, not to say gullible. In this section we discuss these philosophical and psychological claims and suggest that for trust to play the fundamental role it does, it has to be buttressed by active epistemic vigilance.

In Classical epistemology, uncritical acceptance of the claims of others was seen as a failure to meet rationality requirements imposed on genuine knowledge. It did not present the same warrants as clear and distinct ideas or sense impressions arrived at by oneself. According to John Locke, for instance, 'The floating of other men's opinions in our brains makes us not one jot the more knowing, though they happen to be true' (Locke, 1690, book I, ch. 3, sect. 23). Historically, this individualistic stance should be seen as a reaction against the pervasive role in Scholasticism of arguments from authority. It persists in contemporary epistemology, where a common view, described by Tony Coady (1992) as 'reductivist' and by Elisabeth Fricker (1995) as 'reductionist', is that true beliefs acquired through testimony qualify as knowledge only if acceptance of the testimony is itself justified by other true beliefs acquired not through testimony but through perception or inference (see Fricker, 1995; Adler, 2002; van Cleve, 2006).

This reductionist view contrasts with an alternative 'anti-reductionist' approach which treats trust in testimony as intrinsically justified (Hardwig, 1985; Coady, 1992; Foley, 1994). According to Thomas Reid, who provided an early and influential articulation of this anti-reductionist view, humans not only trust what others tell them, but are also entitled to do so. They have been endowed by God with a disposition to speak the truth and a disposition to accept what other people tell them as true. Reid talks of two principles 'that tally with each other,' the *Principle of Veracity* and the *Principle of Credulity* (Reid, 1764, § 24).

Modern defences of a Reidian epistemology appeal to the existence of natural language as material proof that principles of credulity and veracity are indeed in force. How could shared meanings in a public language ever have stabilised, were it not for the fact that most statements in such a language are true testimonials? According to Lewis (1969) and Davidson (1984) in particular, the very possibility of a common language presupposes a generally truthful use of speech. This can be used to provide an a priori justification for trust in testimony (see Coady, 1992). Thus, Tyler Burge argues that linguistic communication has a 'purely preservative character:' just as memory is a medium of content preservation within individuals, so language is a medium of content preservation across individuals. In his view, every act of communication implies a tacit commitment to an acceptance principle which entitles us to 'accept as true something that is presented as true and that is intelligible [to us] unless there are stronger reasons not to do so' (Burge, 1993, pp. 457-88). Approaching the issue from an evolutionary perspective, Ruth Millikan (1987) argues that testimonial communication is a form of perception by proxy and, as such, is a

direct source of knowledge which is no more in need of inferential justification than is knowledge gained from perception.[3]

This debate between reductionism and anti-reductionism revolves around two distinct issues, one normative and the other descriptive. The normative issue has to do with the conditions in which a belief acquired through testimony qualifies as knowledge. The descriptive issue has to do with the cognitive and social practices involved in the production and acceptance of testimony. The two issues are explicitly linked in a 'third way' approach which assumes that our actual practices, which involve some degree of vigilance, are likely to be reasonable, and therefore at least indicative of what the norm should be (e.g. Adler, 2003, Fricker, 2006).

The descriptive issue has recently been taken up in experimental psychology. In particular, work by Daniel Gilbert and his colleagues seems to show that our mental systems start by automatically accepting communicated information, before examining it and possibly rejecting it (Gilbert *et al.,* 1990; Gilbert *et al.*, 1993). This can be seen as weighing (from a descriptive rather than a normative point of view) in favour of an anti-reductionist approach to testimonial knowledge. In a representative experiment, participants were told that they would have to learn Hopi words. They were then presented with sentences such as 'A Monishna is a star', followed shortly by the signal TRUE or FALSE, to indicate the truth value of the preceding statement. In

[3] Note that this anti-reductionist view, which treats testimony as a simple process of content transfer and interpreters as mere receivers of contents, largely ignores or denies the systematically context-dependent and constructive nature of comprehension, which has been well established in modern pragmatics. Bezuidenhout (1998) argues against Burge on this basis, and Origgi and Sperber (2000) criticize Millikan on similar grounds.

some cases, however, participants were distracted while processing the signal TRUE or FALSE. Later, they were given a recognition task in which the same statements about Hopi words were presented, and they had to judge whether they were true or false. The authors predicted that, if acceptance is automatic, as they hypothesised, then distracting participants from indications that the statement was false should lead them to remember it as true. The results confirmed this prediction.

How compelling is Gilbert *et al.*'s evidence? Epistemic vigilance involves a processing cost which is likely to be kept to a bare minimum when the information communicated is of no possible relevance to oneself. So, for instance, if you happen to hear a comment on the radio about a competition in some sport you neither know nor care about, you are unlikely to invest any extra energy in deciding whether or not to believe what you hear. If forced to guess whether it is true or false, you might guess that it is true. After all, it was not merely uttered but asserted. Guessing that it was false would amount to questioning the legitimacy of the assertion, and why should you bother in the circumstances?

More recent experiments have highlighted the crucial role of relevance – or rather, irrelevance – in the materials used by Gilbert and his colleagues. Even if the participants could muster some interest for statements about the meaning of Hopi words (and there is nothing in either the experimental situation or the participants' background knowledge which makes it likely that they would), the information that one of these statements (e.g. 'A Monishna is a star') is false would still be utterly irrelevant to them. From the knowledge that such a statement is false, nothing follows. With other statements, things may be different. If you had prior reasons for thinking that a certain statement was true, or if it described a normal state of affairs, it is easy to see how you might find it relevant to be told that it is false. For instance, it

is easy to see how being told that 'Patrick is a good father' is false might have a wide range of stereotypical implications for you. And indeed, in experiments by Hasson, Simmons, and Todorov (2005) which were otherwise similar to Gilbert's, when participants were presented with statements whose falsity had this kind of potential relevance, automatic acceptance was again no longer found. These results cast doubt on the import of experimental evidence which has been claimed to show that communicated information is automatically accepted (see also Bergstrom and Boyer, submitted; Richter *et al.*, 2009).

As noted above, philosophers and psychologists who argue that humans are fundamentally trustful do not deny that, when the circumstances seem to call for it, people take a critical stance towards communicated information, and may end up rejecting it. So defenders of this approach are not committed to denying that such a critical stance might exploit dedicated cognitive mechanisms for epistemic vigilance. Vigilance (unlike distrust) is not the opposite of trust; it is the opposite of blind trust (see also Yamagishi, 2001). Still, the philosophers and psychologists whose claims we have discussed in this section assume that even if people do not trust blindly, they at least have their eyes closed most of the time to the possibility of being misinformed. In Gilbert's terms, people are trustful 'by default' (Gilbert *et al.,* 1990, p. 601) and are disposed to critically examine communicated information only when circumstances motivate them to do so. This leaves unanswered the question of how they might recognise such circumstances without being vigilant in the first place.

Note too that the idea of default trust draws on an old-style Artificial Intelligence or sequential flow-chart view of cognition, where a mechanism is wholly inactive until its turn comes to do its job, which it then does fully and uninterrupted. An alternative possible view is that several mechanisms may work in parallel or in competition. For

instance, it could be that any piece of communicative behaviour activates two distinct processes in the addressee: one geared to identifying the relevance of what is communicated on the assumption that it is trustworthy, and the other geared to assessing its trustworthiness. Either process might abort for lack of adequate input, or because one process inhibits the other, or as a result of distraction. More generally, acknowledging the paramount importance of trust in human communication need not lead to denying or downplaying the importance of epistemic vigilance.

Here is an analogy which may help to clarify how epistemic trust can co-exist with epistemic vigilance, and indeed be buttressed by it. When we walk down a street through a crowd of people, many at very close quarters, there is a constant risk of inadvertent or even intentional collision. Still, we trust people in the street, and have no hesitation about walking among them. Nor is it just a matter of expecting others to take care while we ourselves walk carelessly. We monitor the trajectory of others, and keep an eye out for the occasional absentminded or aggressive individual, automatically adjusting our level of vigilance to the surroundings. Most of the time, it is low enough to be unconscious and not to detract, say, from the pleasure of a stroll, but it rises when the situation requires. Our mutual trust in the street is largely based on our mutual vigilance. Similarly, in communication, it is not that we can generally be trustful and therefore need to be vigilant only in rare and special circumstances. We could not be mutually trustful *unless* we were mutually vigilant.

3. Comprehension and Acceptance

Human communication is characterised, among other things, by the fact that communicators have two distinct goals: to be understood, and to make their audience think or act according to what is to be understood. Correspondingly, addressees can understand a message without accepting it (whether or not there is a bias or tendency towards acceptance).

Are comprehension and acceptance ever distinct processes in other animals? There is some limited suggestive evidence that they are. It has been experimentally established that vervet and rhesus monkeys do not act on an alarm call from an individual that has produced a series of false alarms calls in the past (Cheney and Seyfarth, 1990; Gouzoules *et al.*, 1996). Interpreted anthropomorphically, this may seem to suggest that the monkeys understand the message, but do not accept it given the unreliability of the source. However, a more parsimonious explanation is that alarm calls from this unreliable individual are not interpreted and then rejected by its conspecifics, but are simply treated as mere noise, and therefore ignored. More generally, there is no strong evidence or argument for distinguishing comprehension from acceptance in non-human communication. (In any case, if it emerged that other social animals do exert some form of epistemic vigilance, this would enrich our understanding of their minds rather than impoverishing our understanding of our own.)

Philosophers of language and pragmatic theorists in the tradition of Austin, Grice and Strawson have been particularly concerned with distinguishing comprehension from acceptance and considering the relations between them. Austin, for instance, distinguished ‘the securing of uptake’ (that is, ‘bringing about the understanding of

the meaning and the force of the locution') from a range of further cognitive or behavioural effects on an audience that he described as 'perlocutionary' (Austin, 1962, p. 116).

Grice (1957) took the speaker's intention to achieve a certain cognitive or behavioural effect that goes beyond the mere securing of uptake as the starting point for his analysis of 'speaker's meaning':

> "[*S*] meant something by *x*" is (roughly) equivalent to "[*S*] intended the utterance of *x* to produce some effect in an audience by means of the recognition of this intention". (Grice 1957/89: 220)

This analysis, which went through a great many revisions and reformulations (e.g. Strawson, 1964; Grice, 1969; Searle, 1969; Schiffer, 1972), treats a speaker's meaning as a complex mental state made up of several layered intentions, of which the most deeply embedded is the intention to make the addressee think or act in a certain way. Beyond this basic intention are two higher-order intentions: that the addressee should recognise the basic intention, and that the addressee's recognition of the basic intention should be at least part of his reason for fulfilling it. By recognising the basic intention (and thus fulfilling the speaker's higher-order intention to have that basic intention recognised), the addressee will have understood the utterance, whether or not he goes on to fulfil the basic intention by producing the desired response.

Following this suggestion about the relation between comprehension and acceptance, Sperber and Wilson (1995) build their inferential model of communication around the idea that speakers have both an *informative* intention

and a *communicative* intention. In their framework, the communicator produces an utterance (or other ostensive stimulus), intending thereby:

> *The informative intention*: to inform the audience of something.
> *The communicative intention:* to inform the audience of one's informative intention.

Here, the informative intention corresponds to Grice's basic-level intention to produce a certain response in an audience, and the communicative intention corresponds to Grice's second-level intention to have this basic intention recognised. Notice that the communicative intention is itself a second-order informative intention, which is fulfilled once the first-order informative intention is recognised. If the addressee accepts the (epistemic or practical) authority of the communicator, recognition of the informative intention will lead to its fulfilment, and hence to the production of the appropriate cognitive or behavioural response. However, the communicative intention can be fulfilled without the corresponding informative intention being fulfilled: in other words, an audience can correctly understand an utterance without accepting or complying with what they have understood.

Sperber and Wilson's analysis of communication departs from Grice's analysis of speaker's meaning in two ways. They argue that only two hierarchically related informative intentions are involved, with the communicative intention being a higher-order intention to inform the audience of one's lower level informative intention. They reject the idea that the communicator must have a third-level intention that the addressee's recognition of her informative intention should be at least part of his reason for fulfilling it.

For Grice, this third-level intention is essential for distinguishing 'meaning' from 'showing'. If I show you that I have a seashell in my pocket, your reason for believing that I have a seashell in my pocket is that you have seen it there. If I tell you "I have a seashell in my pocket," your reason for believing that I have a seashell in my pocket is that I have told you so. How does my telling you something give you a reason to believe it? By the very act of making an assertion, the communicator indicates that she is committing herself to providing the addressee with genuine information, and she intends his recognition of this commitment to give him a motive for accepting a content that he would not otherwise have sufficient reasons to accept. In other words, making an assertion typically involves claiming enough epistemic authority to expect epistemic trust from the addressee. Similarly, making a request typically involves claiming sufficient practical or moral authority to expect the addressee to comply with the request.

But still, is the audience's recognition that the speaker intends her utterance to elicit trust or compliance an intrinsic property of the communication of meaning? Grice himself drew attention to a relevant counter-example which he saw as presenting a problem for his analysis, although he simply mentioned it passing and did not offer any solution. When the communicator is producing a logical argument, she typically intends her audience to accept the conclusion of this argument not on her authority, but because it follows from the premises:

> *Conclusion of argument: p, q, therefore r (from already stated premises)*: While *U[tterer]* intends that *A[ddressee]* should think that *r*, he does not expect (and so intend) *A* to reach a belief that *r* on the basis of *U*'s intention

> that he should reach it. The premises, not trust in *U*, are supposed to do the work. (Grice 1969/1989, p. 107).

Despite the existence of such counter-examples, Grice thought he had compelling reasons to retain this third-level intention in his analysis of 'speaker's meaning'. Sperber and Wilson, on the other hand, were analysing not 'meaning' but 'communication', and they argued that this involves a continuum of cases between 'meaning' and 'showing' which makes the search for a sharp demarcation otiose. In producing an explicit argument, for instance, the speaker both means and shows that her conclusion follows from her premises. Although Grice's discussion of this example was inconclusive, it is relevant to the study of epistemic vigilance. It underscores the contrast between cases where a speaker intends the addressee to accept what she says because she is saying it, and those where she expects him to accept what she says because he recognises it as sound. We will shortly elaborate on this distinction between vigilance towards the source of communicated information and vigilance towards its content.

Clearly, comprehension of the content communicated by an utterance is a precondition for its acceptance. However, it does not follow that the two processes occur sequentially. Indeed, it is generally assumed that considerations of acceptability play a crucial role in the comprehension process itself. We believe that they do, although not in the way commonly envisaged. As noted above, many philosophers have argued that for comprehension to be possible at all, most utterances should be acceptable as true when properly understood. According to Davidson, for instance, we must interpret an utterance 'in a way that optimizes agreement' and that reveals 'a set of beliefs largely consistent and true by our own standard (Davidson, 1984, p. 137).

Such 'interpretive charity' implies an unwillingness to revise one's own beliefs in the light of what others say. It is an a priori policy of trusting others to mean something true – but this is a niggardly form of trust, since it is left up to the interpreter to decide what is true (however, see Davidson, 1986 for a more nuanced picture).

There is a difference between trusting a speaker because you interpret what she says so as to make it as believable to you as possible, and believing what you understand a speaker to say – even if it is incompatible with your own beliefs, which you may then have to revise – because you trust her to start with. In this latter case, interpretation is not guided by a presumption of truth, so what is it guided by? According to Sperber and Wilson (1995), it is guided by an expectation of relevance.

Even when an utterance is in your own language, decoding its linguistic sense falls well short of uniquely determining its interpretation. The comprehension process takes this linguistic sense, together with contextual information, and aims for an interpretation consistent with the expectation of relevance that every utterance elicits about itself. According to relevance theory – we are simplifying here –, every utterance conveys a presumption that it is relevant enough to be worth the hearer's attention. The hearer does not have to accept this presumption: after all, the speaker may not know what is relevant to him, or she may not really care. But whether or not the hearer accepts the presumption of relevance, the very fact that it is conveyed is enough to guide the interpretation process. It justifies the search for an interpretation that the speaker had reason to think would seem relevant to the hearer. In many cases, the output of such a relevance-based interpretation process differs from the one that interpretive charity would select.

To illustrate, suppose Barbara has asked Joan to bring a bottle of champagne to the dinner party:

Andy (to Barbara): A bottle of champagne? But champagne is expensive!

Barbara: Joan has money.

Andy had previously assumed that Joan was just an underpaid junior academic. How should he interpret Barbara's reply? If his aim was to optimize agreement, he should take Barbara to be asserting that Joan has some money, as opposed to no money at all, which is true of most people and which he already believes is true of Joan. But interpreted in this way, Barbara's reply is not relevant enough to be worth Andy's attention. By contrast, if he interprets 'has money' as intended to convey *has enough money to be easily able to afford champagne* Barbara's utterance would be relevant enough. More precisely, it would be relevant enough to Andy provided that he believes it. If he does not believe what he takes Barbara to say, then her utterance will only provide him with information about Barbara herself (her beliefs, her intended meaning) rather than about Joan, and this may not be relevant enough to him. How, then, do considerations of relevance help Andy understand Barbara's meaning when he is not willing to accept it? We claim that, whether he ends up accepting it or not, the hearer interprets the speaker as asserting a proposition that would be relevant enough to him provided that he accepted it.

In other words, hearers must adopt a 'stance of trust' in the course of interpretation (see Holton, 1994; Origgi, 2005, 2008b). The trust required is less miserly than what is required by the principle of charity. It involves a readiness to adjust one's own beliefs to a relevance-guided interpretation of the speaker's meaning, as opposed to adjusting one's interpretation of the speaker's meaning to one's own beliefs. On the other hand, it is tentative trust. We claim that interpreting

an utterance as if one were going to accept it is not tantamount to actually accepting it, not even to accepting it by 'default'.

Still, it might be that the stance of trust involved in comprehension causes, or contributes to causing, a tendency or bias in favour of actual acceptance of communicated information. If so, this might help to explain the results of Gilbert's psychological experiments and the introspective considerations that motivate Reidian philosophers to assume (wrongly) that epistemic trust is a default disposition. Note that a mere tendency or bias in favour of accepting communicated information would not be irrational since, presumably, most communication is honest (and is so, we maintain, in at least partly because the audience's vigilance limits the range of situations where dishonesty might be in the communicators' best interest).

So, understanding is not believing, but nor is it adopting a sceptical position. Comprehension involves adopting a tentative and labile stance of trust; this will lead to acceptance only if epistemic vigilance, which is triggered by the same communicative acts that trigger comprehension, does not come up with reasons to doubt.

4. Vigilance towards the Source

Communication brings vital benefits, but carries a major risk for the audience of being accidentally or intentionally misinformed. Nor is there any failsafe way of calibrating one's trust in communicated information so as to weed out all and only the misinformation. Given that the stakes are so high, it is plausible that there has been ongoing selective pressure in favour of any available cost-effective means to

least approximate such sorting. Since there are a variety of considerations relevant to the granting or withholding of epistemic trust, we will explore the possibility that different abilities for epistemic vigilance may have emerged in biological and cultural evolution, each specialising in a particular kind of relevant considerations.

Factors affecting the acceptance or rejection of a piece of communicated information may have to do either with the source of the information – who to believe – or with its content – what to believe. In this section and the next, we consider epistemic vigilance directed at the source of information.

Judgements about the trustworthiness of informants may be more or less general or contextualised. You may think, 'Mary is a trustworthy person,' meaning it both epistemically and morally, and therefore expecting what Mary says to be true, what she does to be good, and so on. Or you may trust (or mistrust) someone on a particular topic in specific circumstances: 'You can generally trust Joan on Japanese prints, but less so when she is selling one herself.' Trust can be allocated in both these ways, but how do they compare from a normative point of view?

A reliable informant must meet two conditions: she must be competent, and she must be benevolent. That is, she must possess genuine information (as opposed to misinformation or no information), and she must intend to share that genuine information with her audience (as opposed to making assertions she does not regard as true, through either indifference or malevolence). Clearly, the same informant may be competent on one topic but not on others, and benevolent towards one audience in certain circumstances, but not to another audience or in other circumstances. This suggests that trust should be allocated to informants depending on the topic, the audience, and the circumstances. However such precise calibration of trust is costly in cognitive terms, and, while people are often willing to pay the price, they also

commonly rely on less costly general impressions of competence, benevolence and overall trustworthiness.

A striking illustration of the tendency to form general judgments of trustworthiness on the basis of very limited evidence is provided in a study by Willis and Todorov (2006). Participants were shown pictures of faces, for either a mere 100 milliseconds or with no time limit, and asked to evaluate the person's trustworthiness, competence, likeability, aggressiveness and attractiveness. Contrary to the authors' expectations, the correlation between judgments with and without time limit was not greater for attractiveness (.69) – which is, after all, a property of a person's appearance – than for trustworthiness (.73), while the correlations for aggressiveness and competence were a relatively low .52. One might wonder if such split-second judgments of trustworthiness have any basis at all, but what this experiment strongly suggests is that looking for signs of trustworthiness is one of the first things we do when we see a new face (see also Ybarra *et al.*, 2001).

There is a considerable social psychology literature suggesting that people's behaviour is determined to a significant extent not by their character but by the situation (Ross and Nisbett, 1991; Gilbert and Malone, 1995). If so, judging that someone is generally trustworthy may be a case of the 'fundamental attribution error' (Ross, 1977): that is, the tendency, in explaining or predicting someone's behaviour, to overestimate the role of psychological dispositions and underestimate situational factors. But even without appealing to character psychology, it is possible to defend the view that some people are more generally trustworthy than others, and are to some extent recognisable as such.

If we continually interact with the same people, misinforming them when it is to our own immediate advantage may damage our reputation and end up being costly in

the long run. Conversely, doing our best to be systematically trustworthy may sometimes be costly in the short run, but may be beneficial in the long run. The trade-off between the short term cost and long term benefits of a policy of trustworthiness may differ from person to person, depending, for instance, on the way they discount time (Ainslie, 2001), and they may end up following different policies. If such policies exist, then general judgments of relative trustworthiness might not be baseless. People who opt for a policy of systematic trustworthiness would stand to benefit from a reputation for being highly trustworthy. This reputation would be fed by common knowledge of their past actions, and might be further advertised by their everyday public behaviour and demeanour.

It is possible to project an image of trustworthiness (whether or not that image is itself trustworthy). Is it also possible, conversely, to engage in deceptive behaviour such as lying without giving any detectable evidence of the fact? There is a substantial literature on lie detection (see Ekman, 2001, for a review), and what it shows, in a nutshell, is that detecting lies on the basis of non-verbal behavioural signs is hard (Vrij, 2000; Malone and DePaulo, 2001; Bond and DePaulo, 2006), even for people who are trained to do so (e.g. DePaulo and Pfeifer, 1986; Ekman and O'Sullivan, 1991; Mann, Vrij and Bull, 2004; Vrij, 2004), and even when the liars are far from expert – for instance, when they are three-year-old children (Lewis *et al.*, 1989; Talwar and Lee, 2002). The ability to lie can be quite advantageous, but only if the liars do not give themselves away. Whatever the respective contributions of evolved dispositions and acquired skills, liars seem able to keep the behavioural signs of dishonesty to a minimum.

In order to gain a better grasp of the mechanisms for epistemic vigilance towards the source, what is most urgently needed is not more empirical work on lie detection

or general judgments of trustworthiness, but research on how trust and mistrust are calibrated to the situation, the interlocutors and the topic of communication. Here, two distinct types of consideration should be taken into account: the communicator's competence on the topic of her assertions, and her motivation for communicating. Both competence and honesty are conditions for believability. There is a considerable literature with some indirect relevance to the study of epistemic vigilance in ordinary communication, for instance in the history and sociology of science (e.g. Shapin, 1994), the anthropology of law (e.g. Hutchins, 1980; Rosen, 1989), the linguistic study of evidentials (e.g. Chafe and Nichols, 1986; Ifantidou, 2001; Aikhenvald, 2004), or the social psychology of influence and persuasion (e.g. Chaiken, 1980; Petty and Cacioppo, 1986). However, much more work needs to be done on epistemic vigilance in everyday communication.

In the next section, we turn to the development of vigilance towards the source in childhood, which is not only interesting in its own right, but will also help us separate out the various components of epistemic vigilance towards the source of information.

5. The Development of Epistemic Vigilance (and Mindreading)

There is a growing body of research on the development of children's epistemic vigilance (for reviews, see e.g. Koenig and Harris, 2007; Heyman, 2008; Clément, In press; Corriveau and Harris, In press; Nurmsoo *et al.*, In press). This shows that even at a very early age, children do not treat all communicated information as equally reliable. At 16 months, they notice when a familiar word is inappropriately used (Koenig and Echols, 2003). By the age of two, they often attempt to contradict and

correct assertions that they believe to be false (e.g. Pea, 1982). These studies challenge the widespread assumption that young children are simply gullible.

Do young children have the cognitive resources to allocate trust on the basis of relevant evidence about an informant's trustworthiness? Given the choice, three-year-olds seem to prefer informants who are both benevolent (Mascaro and Sperber, 2009) and competent (e.g. Clément *et al.*, 2004). In preferring benevolent informants, they take into account not only their own observations but also what they have been told about the informant's moral character (Mascaro and Sperber, 2009), and in preferring competent informants, they take past accuracy into account (e.g. Clément *et al.*, 2004; Birch *et al.*, 2008; Scofield and Behrend, 2008). By the age of four, they not only have appropriate preferences for reliable informants, but also show some grasp of what this reliability involves. For instance, they can predict that a dishonest informant will provide false information (Couillard and Woodward, 1999), or that an incompetent informant will be less reliable (Call and Tomasello, 1999; Lampinen and Smith, 1995; Clément et al., 2004). Moreover, they make such predictions despite the fact that unreliable informants typically present themselves as benevolent and competent.

Early epistemic vigilance draws on some of the capacities used in selecting partners for cooperation, which include moral evaluation, monitoring of reliability, and vigilance towards cheating (e.g. Cosmides and Tooby, 2005, Harris and Núñez, 1996). Indeed, the exercise of epistemic vigilance not only relies on some of the processes involved in selecting cooperative partners, but also contributes to their success. In particular, it contributes to the relative reliability of reputation systems, a fundamental tool for selecting cooperative partners (e.g. Alexander, 1987; Nowak and Sigmund, 1998; Milinski *et al.*, 2002 – but see below). For instance, as four- and five-

year-old children become increasingly vigilant towards deception (Couillard and Woodward, 1999; Mascaro and Sperber, 2009) they also become more vigilant towards hypocrisy in self-presentation (Peskin, 1996; Mills and Keil, 2005; Gee and Heyman, 2007).

Epistemic vigilance directed at informants yields a variety of epistemic attitudes (acceptance, doubt or rejection, for instance) to the contents communicated by these informants. There is some evidence that three-year-old children are aware of attitudes such as endorsement or doubt (Fusaro and Harris, 2008), and are also aware that assertions can be stronger or weaker (Sabbagh and Baldwin, 2001; Birch *et al.*, 2008; Matsui *et al.,* 2009). Children are able to make sense of comments on the reliability of what is communicated (e.g. Fusaro and Harris, 2008, Clément *et al.,* 2004). As a result, they can take advantage of the epistemic judgments of others, and enrich their own epistemological understanding and capacity for epistemic vigilance in doing so.

Children also appear to have some capacity to compare the reliability of different sources of information. In experiments modelled on Solomon Asch's famous studies on conformity (Asch, 1956), for instance, a majority of three-year-olds trust their own perceptions rather than a series of consistently false judgments made by confederates of the experimenters (Corriveau and Harris, In press, although see Walker and Andrade, 1996). Children take account of an informant's access to information (e.g. Robinson and Whitcombe, 2003; Robinson *et al.*, 2008; (Nurmsoo and Robinson, 2009). They also attribute to others lasting dispositions for greater or lesser reliability (e.g. Koenig and Harris, 2007; Birch *et al.*, 2009; Corriveau and Harris, 2009), and may do this on the basis of an understanding that different people are more or less knowledgeable – a component of the child's naïve psychology which has not been

investigated in depth (though see Lutz and Keil, 2002). Children's epistemic vigilance thus draws on – and provides evidence for – distinct aspects of their naïve epistemology: their understanding that people's access to information, strength of belief, knowledgeability, and commitment to assertions come in degrees.

When epistemic vigilance is targeted at the risk of deception, it requires an understanding not only of a communicator's epistemic states but also of her intentions, including intentions to induce false beliefs in her audience. This calls for relatively sophisticated mindreading using higher-order metarepresentations ('She believes that not-P but wants me to believe that P' combines a first-order attribution of belief with a second-order attribution of intention).

There are interesting parallels between the development of epistemic vigilance and evidence from false belief tasks classically used to measure the development of mindreading. Rudiments of epistemic vigilance are found in early childhood, arguably even in infancy. However, starting at around the age of four, there is a major transition in children's epistemic vigilance towards both dishonesty (Mascaro and Sperber, 2009; see also Couillard and Woodward, 1999; Jaswal *et al.* In press) and incompetence (Povinelli and DeBlois, 1992; Call and Tomasello, 1999; Welch-Ross, 1999; Figueras-Costa and Harris, 2001). At four, children begin to show increased attention to the epistemic quality of other people's beliefs and messages. They become much more selective in their trust, and also much more willing and able to manipulate the beliefs of others.

This transition in epistemic vigilance occurs around the age at which children succeed in passing standard false belief tasks (Wimmer and Perner, 1983; Baron-Cohen *et al.*, 1985). Until recently, this convergence might have been interpreted on the following lines: At around four years of age, as a result of their emerging

understanding of belief – and false belief in particular – children become increasingly aware that others may hold false beliefs, and (at a higher metapresentational level) that others may want them to hold false beliefs. This awareness is the basis for more adult-like forms of epistemic vigilance. Whatever the attractions of this line of interpretation, it has become less plausible as a result of recent experiments with new non-verbal versions of the false belief task adapted for use with infants. These experiments suggest that by their second year, children already expect an agent's behaviour to be guided by its beliefs, even when they are false (e.g. Onishi and Baillargeon, 2005; Southgate *et al.*, 2007; Surian *et al.*, 2007). If so, the robust results of work with the standard false belief task must be reinterpreted, and so must the transition that takes place around the age of four.

Two possible reinterpretations of this transition come readily to mind (and a full picture might draw on both). First, the ability to pass standard false belief tasks and the improved capacity for epistemic vigilance might have a common cause, for instance, a major development in executive function abilities (e.g. Perner and Lang, 2000; Carlson and Moses, 2001). Second, as a result of their improved capacity for epistemic vigilance, children may start paying attention to relevant aspects of false belief tasks which are generally missed at an earlier age. As they become increasingly aware that others may hold false beliefs (through either epistemic bad luck or deception), they get better at taking these false beliefs into account when predicting the behaviour of others. Their interest here is not so much that of an observer, but rather that of a potential victim of misinformation, a potential perpetrator of deception, or a co-operator who prefers knowledgeable partners.

Current studies of epistemic vigilance thus offer some interesting insights into the nature and development of theory of mind abilities. They show that epistemic

vigilance draws on a variety of cognitive mechanisms with distinct developmental trajectories, including the moral sense involved in recognising potential partners for cooperation, naïve epistemology, and mindreading.

6. Vigilance towards the Content

As we have seen in the last two sections, epistemic vigilance can be directed at the source of communicated information: Is the communicator competent and honest? It can also be directed at the content of communication, which may be more or less believable independently of its source. In this section and the next, we consider epistemic vigilance directed at the content of communication.

Some contents are intrinsically believable even if they come from an untrustworthy source. Examples include tautologies, logical proofs, truisms, and contents whose truth is sufficiently evidenced by the act of communication itself (e.g. saying, 'Je suis capable de dire quelques mots en français'). Other contents are intrinsically unbelievable even if they come from a trustworthy source. Examples include logical contradictions, blatant falsehoods, and contents whose falsity is sufficiently evidenced by the act of communication itself (e.g. saying, 'I am mute').

In most cases, however, epistemic vigilance directed at communicated content must rely on more than just its inherent logical properties, indisputable background knowledge, or the self-confirming or -disconfirming nature of some utterances. The believability of newly communicated information must be assessed relative to background beliefs which are themselves open to revision. Obviously, new information cannot be assessed relative to the whole of one's 'mental encyclopaedia'.

To keep processing time and costs within manageable limits, only a very small subset of that encyclopaedia, closely related to the new piece of information, can be brought to bear on its assessment. Indeed, the systematic activation of even a limited subset of background information solely for the purpose of assessing the believability of communicated content would still be quite costly in processing terms. We will argue that such ad hoc activation is unnecessary.

According to relevance theory (Sperber and Wilson, 1995, 2005; Carston, 2002; Wilson and Sperber, 2004), the comprehension process itself involves the automatic activation of background information in the context of which the utterance may be interpreted as relevant. Here, the processing costs tend to be proportionate to the cognitive benefits derived. We claim that this same background information which is used in the pursuit of relevance can also yield an imperfect but cost-effective epistemic assessment. Moreover, as we will now show, the search for relevance involves inferential steps which provide a basis for this assessment.

Sperber and Wilson (1995) distinguish three types of contextual effect through which a piece of new information can achieve relevance in a context of existing beliefs. (i) When new information and contextual beliefs are taken together as premises, they may yield 'contextual implications' (implications derivable from neither the context nor the new information alone) which are accepted as new beliefs. (ii) The individual's confidence in contextually activated beliefs may be raised or lowered in the light of new information. (iii) New information may contradict contextually activated beliefs and lead to their revision. All three types of contextual effect (acceptance of contextually implied new beliefs, modification of strength of beliefs, and revision of beliefs) tend to contribute to an improvement in the individual's knowledge.

What happens when the result of processing some new piece of information in a context of existing beliefs is a contradiction? When the new information was acquired through perception, it is quite generally sound to trust one's own perceptions more than one's memory and to update one's beliefs accordingly. You believed Joan was in the garden; you hear her talking in the living room. You automatically update your belief about Joan's whereabouts. Presumably, such automatic updating is the only form of belief revision engaged in by non-human animals.

When the new information was communicated, on the other hand, there are three possibilities to consider. (i) If the source is not regarded as reliable, the new information can simply be rejected as untrue, and therefore irrelevant: for instance, a drunk in the street tells you that there is a white elephant around the corner. (ii) If the source is regarded as quite authoritative and the background beliefs which conflict with what the source has told us are not held with much conviction, these beliefs can be directly corrected: for instance, looking at Gil, you had thought he was in his early twenties, but he tells you that he is 29 years old. You accept this as true and relevant – relevant in the first place because it allows you to correct your mistaken beliefs. (iii) If you are confident about both the source and your own beliefs, then some belief revision is unavoidable. You must revise either your background beliefs or your belief that the source is reliable, but it is not immediately clear which. For instance, it seemed to you that Gil was in his early twenties; Lucy tells you that he must be in his early thirties. Should you should stick to your own estimate or trust Lucy's?

Things are not so different when the result of processing information communicated by a trusted source is not a logical inconsistency but an empirical incoherence: that is, when the new information is incompatible with some of our background beliefs, given other more entrenched background beliefs. For instance,

you believed that Gil was a doctor; Lucy tells you that he is only 22 years old. You have the well entrenched belief that becoming a doctor takes many years of study, so that it is almost impossible to be a doctor by the age of 22. Hence, you should either disbelieve Lucy or give up the belief that Gil is a doctor. In order to preserve coherence, you must reduce either your confidence in the source or your confidence in your less entrenched beliefs.

The role of coherence checking in belief revision has been highlighted by philosophers such as Gilbert Harman (1986) and Paul Thagard (2002). Here, though, we see coherence checking not as a general epistemic procedure for belief revision, but as a mechanism for epistemic vigilance directed at communicated content, which takes advantage of the limited background information activated by the comprehension process itself.

If only for reasons of efficiency, one might expect the type of coherence checking used in epistemic vigilance to involve no more than the minimal revisions needed to re-establish coherence. In some cases, coherence is more easily restored by distrusting the source, and in others by revisiting some of one's own background beliefs. Unless one option dominates the competition to the point of inhibiting awareness of the alternatives, it takes a typically conscious decision to resolve the issue. Making such a decision involves engaging in some higher order or metapreresentational thinking about one's own beliefs.

What we are suggesting is that the search for a relevant interpretation, which is part and parcel of the comprehension process, automatically involves the making of inferences which may turn up inconsistencies or incoherences relevant to epistemic assessment. When such inconsistencies or incoherences occur, they trigger a procedure wholly dedicated to such assessment. Still, comprehension, the search for

relevance, and epistemic assessment are interconnected aspects of a single overall process whose goal is to make the best of communicated information.

7. Epistemic Vigilance and Reasoning

Now consider things from the communicator's point of view. Suppose she suspects that her addressee is unlikely to accept what she says purely on trust, but will probably exercise some epistemic vigilance and check how far her claim coheres with his own beliefs. The addressee's active vigilance stands in the way of the communicator's achieving her goal. Still, from the communicator's point of view, a vigilant addressee is better than one who rejects her testimony outright. And indeed, the addressee's reliance on coherence as a criterion for accepting or rejecting her claim may offer the communicator an opportunity to get past his defences and convince him after all.

We have suggested that coherence checking takes place against the narrow context of beliefs used in the search for a relevant interpretation of the utterance. But the addressee may have other less highly activated beliefs which would have weighed in favour of the information he is reluctant to accept, if he had been able to take them into account. In that case, it may be worth the communicator's while to remind the addressee of these background beliefs, thus increasing the acceptability of her claim. Or there may be other information that the addressee would accept on trust from the communicator, which would cohere well with her claim and thus make it more acceptable.

To illustrate, we will adapt a famous example from Grice (1975/1989, p. 32). Andy and Barbara are in Boston, gossiping about their friend Steve:

Andy: Steve doesn't seem to have a girlfriend these days

Barbara: He has been paying a lot of visits to New York lately

Barbara believes that Steve has a new girlfriend, but feels that if she were simply to say so, Andy, who has just expressed doubt on the matter, would disagree. Still, she has noticed that Steve has been paying a lot of visits to New York lately, and regards this as evidence that Steve has a girlfriend there. Andy might also have noticed these visits, and if not, he is likely to accept Barbara's word for it. Once he takes these visits into account, the conclusion that Steve might have a girlfriend may become much more acceptable to him.

Barbara is making no secret of the fact that she wants Andy to accept this conclusion. On the contrary, her assertion that Steve has been paying a lot of visits to New York will only satisfy Andy's expectations of relevance (or be cooperative in Grice's sense) to the extent that it is understood as implicating that Steve might have a girlfriend despite Andy's doubts. Although Andy recognises this implicature as part of Barbara's intended meaning, he may not accept it. What Barbara is relying on in order to convince him is not his ability to understand her utterance but his ability to grasp the force of the argument whose premises include her explicit statement, together with other pieces of background knowledge (about Steve's likely reasons for visiting New York regularly), and whose conclusion is her implicature.

In another, slightly different scenario, Andy himself remarks that Steve has been paying a lot of visits to New York lately; but failing to see the connection, he adds, 'He

doesn't seem to have a girlfriend these days.' In that case, Barbara may highlight the connection in order to help him come to the intended conclusion, saying: 'If he goes to New York, it may be to see a girlfriend', or 'If he had a girlfriend in New York, that would explain his visits.' Or she might simply repeat Andy's comment, but with a different emphasis: 'He doesn't *seem* to have a girlfriend these days, but he *has* been paying a lot of visits to New York lately.' Logical connectives such as 'if' and discourse connectives such as 'but', which suggest directions for inference, are used by the communicator to help the addressee arrive at the intended conclusion (Blakemore, 1987, 2002).

In both scenarios, when Barbara expresses herself as she does and Andy sees the force of her implicit argument, they are making use of an inferential mechanism which is sensitive to logical and evidential relationships among propositions, and which recognises, more specifically, that some function as premises and others as conclusions. What Barbara conveys, and what Andy is likely to recognise, is that it would be incoherent to accept the premises and reject the conclusion.

Argumentation, in either the simple and largely implicit form illustrated in the above scenarios or in more complex and more explicit forms, is a product of reasoning.[4] In a series of papers (Mercier and Sperber, 2009; Sperber and Mercier, In

[4] The term 'reasoning' is sometimes used in a broad sense as a synonym of 'inference' (in particular in developmental and comparative psychology). Here we use 'reasoning' in its more frequent sense to refer to a form of inference which involves attending to the reasons for accepting some conclusion. Reasoning, so understood, involves reflection, and contrasts with intuitive forms of inference where we arrive at a conclusion without attending to reasons for accepting it. A similar contrast between intuitive and reflective forms of inference has been much discussed under the

press; Sperber, 2001; Mercier and Sperber, Forthcoming) Hugo Mercier and Dan Sperber have argued that reasoning is a tool for epistemic vigilance, and for communication with vigilant addressees. Its main function is to enable communicators to produce arguments designed to convince others, and addressees to evaluate arguments so as to be convinced only when appropriate.

Classically, reasoning is seen as a tool for individual cognition, which is supposed to help people overcome the limits of intuition, acquire better grounded beliefs, particularly in areas beyond the reach of perception and spontaneous inference, and make good decisions (Evans and Over, 1996; Kahneman, 2003; Stanovich, 2004). This common view is not easy to square with the massive evidence that human reasoning is not so good at fulfilling this alleged function. Ordinary reasoning fails to solve trivial logical problems (Evans, 2002) or override transparently flawed intuitions (Denes-Raj and Epstein, 1994). It often leads us towards bad decisions (Shafir *et al.*, 1993; Dijksterhuis *et al.*, 2006) and poor epistemic outcomes (Kunda, 1990). By contrast, intuition has a good track record for efficiently performing very complex computations (e.g. (Trommershauser *et al.*, 2008; Balci *et al.*, 2009). Human reasoning, with its blatant shortcomings and relatively high operating costs, is not properly explained by its alleged function as a tool for individual cognition.

Predictions derived from the 'argumentative theory of reasoning', by contrast, are supported by a wealth of evidence from different fields in psychology (Mercier, submitted-b, submitted-a, submitted-c; Mercier and Landemore, submitted; Mercier and Sperber, submitted). To give just one example, the argumentative theory makes a

---

heading of 'dual process' theories of reasoning (see, for instance, Evans and Frankish, 2009).

prediction which sets it apart from other approaches and which is of particular relevance to the study of epistemic vigilance. If the function of reasoning is to find arguments to convince others, then the arguments it comes up with should support the communicator's position, and, in appropriate circumstances, should undermine the interlocutor's position. In other words, reasoning should exhibit a strong *confirmation bias*. And indeed, this bias towards 'seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand' (Nickerson, 1998 p.175) has been evidenced in countless psychology experiments and observations in natural settings (see Nickerson, 1998 for review). For classical accounts of reasoning, it is puzzling that such a bias should be so robust and prevalent – or indeed that it should exist at all – and attempts have been made to explain it away by appealing to motivational or cognitive limitations. But there are several arguments against these attempted explanations. In the first place, the confirmation bias seems to be restricted to reasoning, and not to occur in intuitive judgments (Mercier and Sperber, Forthcoming). In the second place, attempts to overcome these alleged cognitive or motivational problems make little or no difference (Camerer and Hogarth, 1999; Willingham, 2008). This suggests that the confirmation bias is not a *flaw* in reasoning, but rather a *feature* that is to be expected in a mechanism designed to persuade others by use of arguments.

Interestingly, the confirmation bias need not lead to poor performance from a logical normative point of view. When people with different viewpoints share a genuine interest in reaching the right conclusion, the confirmation bias makes it possible to arrive at an efficient division of cognitive labour. Each individual looks only for reasons to support their own position, while exercising vigilance towards the arguments proposed by others and evaluating them carefully. This requires much less

work than having to search exhaustively for the pros and cons of *every* position present in the group. By contrast, when the confirmation bias is not held in check by others with dissenting opinions, reasoning becomes epistemically hazardous, and may lead individuals to be over-confident of their own beliefs (Koriat *et al.*, 1980), or to adopt stronger version of those beliefs (Tesser, 1978). In group discussions where all the participants share the same viewpoint and are arguing not so much against each other as against absent opponents, such polarization is common and can lead to fanaticism (Sunstein, 2002).

We are not claiming that reasoning takes place only in a communicative context. It clearly occurs in solitary thinking, and plays an important role in belief revision. We would like to speculate, however, that reasoning in non-communicative contexts is an extension of a basic component of the capacity for epistemic vigilance towards communicated information, and that it typically involves an anticipatory or imaginative communicative framing. On this view, the solitary thinker is in fact considering claims she might be presented with, or that she might want to convince others to accept, or engaging in a dialogue with herself where she alternates between different points of view. Experimental evidence might help confirm or disconfirm such speculation: for instance, we predict that encouraging or inhibiting such mental framing would facilitate or hamper reasoning.

8. Epistemic vigilance on a population scale

What we have considered so far is the filtering role that epistemic vigilance plays in the flow of information in face to face interaction. In this section, we turn to the flow

of information on a population scale, as for instance in the emergence of good or bad reputations or the propagation of religious beliefs. Information of this type that spreads in a population through social transmission is known as 'cultural information'. The very social success which is almost a defining feature of cultural information might suggest that (except in cases of cultural conflict) it is uncritically accepted. We will argue, however, that here too epistemic vigilance is at work, but that it needs appropriate cultural and institutional development to meet some of the epistemic challenges presented by cultural information.

No act of communication among humans, even if it is only of local relevance to the interlocutors at the time, is ever totally disconnected from the flow of information in the whole social group. Human communication always carries cultural features. It may do so explicitly, as when Andy says to Barbara, 'Champagne is expensive!', reminding her of an assumption which is culturally shared in their milieu, and which he sees as relevant in the circumstances. It may do so implicitly, as when Andy says to Barbara in another of our examples, 'Steve doesn't seem to have a girlfriend these days.' Although Andy's remark has quite limited and local relevance, it implicates culturally shared assumptions about what is to be expected of a bachelor like Steve, without which his remark would not be relevant in the intended way. Many cultural assumptions are distributed in this way, not so much – or, in some cases, not at all – by being directly asserted, but by being used as implicit premises in a vast number of communicative acts such as Andy's utterance.

Is epistemic vigilance exercised in the case of culturally transmitted contents, and if so, how? When contents of this type are conveyed, either explicitly or implicitly, communicators use their own individual authority not so much to endorse the content as to vouch for its status as a commonly accepted cultural assumption. When

Andy says that champagne is expensive, or implies that it would be normal for Steve to have a girlfriend, he conveys that these are accepted views in his and Barbara's milieu. If Barbara disagrees, her disagreement is not just with Andy, but with this accepted view.

If an idea is generally accepted by the people you interact with, isn't this a good reason for you to accept it too? It may be a modest and prudent policy to go along with the people one interacts with, and to accept the ideas they accept. Anything else may compromise one's cultural competence and social acceptability. For all we know, it may be quite common for members of a cultural group to accept what they take to be 'accepted views' in this pragmatic sense, without making any strong or clear epistemic commitment to their content (Sperber, 1975, Sperber, 1985; Boyer 1992; Bloch 1998). From an epistemological point of view, the fact that an idea is widely shared is not a good reason to accept it unless these people have come to hold it independently of one another. Only in those circumstances will every individual who accepts this idea add to our own epistemological reasons for accepting it too. Quite often, however, people who accept (in an epistemological sense) culturally shared ideas have no independent reasons for doing so.

Often, information spreads through a group from a single source, and is accepted by people along the chains of transmission because they trust the source rather than because of any evidence or arguments for the content. If so, the crucial consideration should be the trustworthiness of the original source. If each person who passes on the information has good independent reasons for trusting the source, this should give people further along the chain good reasons for also trusting the source, and thus for accepting the content originally conveyed. However, people's reasons for trusting the

source are in general no more independent of one another than their reasons for accepting the content.

Even if we make the strong assumption that each individual along the chain of transmission from the source to ourselves had good reasons for trusting the previous individual in the chain, these reasons can never be error-proof; hence, our own confidence in the original source should diminish as the length of the chain increases. Moreover, it is quite common for a piece of information with no clearly identified source to be accepted and transmitted purely on the ground that it is widely accepted and transmitted – an obvious circularity. Add to all this the fact that when an idea propagates through a population, its content tends to alter in the process without the propagators being aware of these alterations (as with nearly all rumours and traditions – see Sperber, 1996). In these cases, even if there were good reason to regard the original source as reliable, this would provide no serious support for the idea as currently formulated.

It might seem, then, that people are simply willing, or even eager, to accept culturally transmitted information without exercising ordinary epistemic vigilance towards it. Boyd, Richerson and Henrich have argued that there is an evolved conformist bias in favour of adopting the behaviour and attitudes of the majority of members of one's community (e.g. (Boyd and Richerson, 1985; Henrich and Boyd, 1998). Csibra and Gergely (2009) have argued that people in general, and children in particular, are eager to acquire cultural information, and that this may bias them towards interpreting (and even over-interpreting) communicated information as having cultural relevance, and also towards accepting it. An alternative (or perhaps complementary) hypothesis is that people do exercise some degree of epistemic vigilance towards all communicated information, whether local or cultural, but that

their vigilance is directed primarily at information originating in face to face interaction, and not at information propagated on a larger scale. For instance, people may be disposed to pay attention to the problems raised by the non-independence of testimonies, or by discrepancies in their contents, when they are blatantly obvious, as they often are when they occur in face to face interaction, but not otherwise. On a population scale, these problems can remain unnoticed although, on reflection, they are likely to be pervasive. All kinds of beliefs widely shared in the community may propagate throughout a culture by appealing to individual trust in converging testimonies. The trust is not blind, but the epistemic vigilance which should buttress it is short-sighted.

Of particular relevance here are two kinds of belief which are typically cultural: reputations, and beliefs which are only partly understood, and whose content is mysterious. There is a vast literature on reputation in general (e.g.. Tirole, 1996; Morris, 1999) and on its relevance to cooperation and to social epistemology (e.g. Mathew and Boyd, 2009).

The term ‘reputation’ is generally understood as a positive or negative opinion, for instance, the opinion that Lisa is generous or that John is a liar, which has become widely accepted in a group through repeated transmission. When an individual belongs to a relatively small group in which many people have direct experience of her qualities and shortcomings, and where they can express and compare their opinions with some freedom (for instance by gossiping), then her actual behaviour may play an important role in reinforcing or compromising her reputation. Of course, gossips may themselves be incompetent or not quite honest, but ordinary epistemic vigilance is relevant to assessing both gossipers and gossip. However, many reputations are spread on a larger scale, by people with no knowledge relevant to

their direct assessment. When an addressee has to decide whether or not to believe an unfamiliar source of information, she may have no other basis for her decision than her knowledge of the source's reputation, which she is unable to assess herself, and which she is likely to accept for want of a better choice. All too often, reputations are examples of ideas which are accepted and transmitted purely on the ground that they are widely accepted and transmitted.

As noted above, the content of socially transmitted beliefs is typically modified in the course of transmission. One of the ways in which reputations get transformed is by becoming inflated well beyond the level found in typical opinions arrived at individually. When epistemic authorities – religious leaders, gurus, *maîtres à penser* – achieve such inflated reputations, people who are then inclined to defer more to them than to any source whose reliability they have directly assessed may find themselves in the following predicament: If they were to check the pronouncements of these sources (for instance, 'Mary was and remained a virgin when she gave birth' or Lacan's 'There is no such thing as a sexual relationship') for coherence with their existing beliefs, they would reject them. But this would in turn bring into question their acceptance of the authority of the source. A common solution to this predicament is to engage in a variant of Davidsonian 'charitable interpretation', and to 'optimize agreement' not by providing a clear and acceptable interpretation of these pronouncements, but by deferring to the authorities (or their authorised interpreters) for the proper interpretation, and thus accepting a half-understood or 'semi-propositional' idea (Sperber, 1985; 1997; In press). Most religious beliefs are typical examples of beliefs of this kind, whose content is in part mysterious to the believers themselves (Bloch, 1998; Boyer 2001).

So far, the picture we have sketched of epistemic vigilance on a population scale is

somewhat grim. Mechanisms for epistemic vigilance are not geared to filtering information transmitted on such a large scale. Even if we are right to claim that these mechanisms exist, they do not prevent mistaken ideas, undeserved reputations and empty creeds from invading whole populations. However, we did note that it is important not to jump from the fact that people are seriously, even passionately, committed to certain ideas, and expect others to be similarly committed, to the conclusion that the commitment involved is clearly *epistemic*. It may be that the content of the ideas matters less to you than who you share them with, since they may help define group identities. When what matters is the sharing, it may be that contents which are unproblematically open to epistemic evaluation would raise objections within the relevant social group, or would be too easily shared beyond that group. So, semi-propositional contents which can be unproblematically accepted by just the relevant group may have a cultural success which is negatively correlated with their epistemic value.

So far, we have considered only the effects of ordinary individual vigilance exercised on a population scale. However, epistemic vigilance can also take on institutional form. Some of these institutions help to protect established authorities or impose a dogma, and are therefore detrimental to true epistemological goals. Strictly speaking, such forms of hegemonic or dogmatic vigilance are not epistemic. By contrast, other institutional forms of genuinely epistemic vigilance may provide better epistemic filtering than the mere cumulative effect of spontaneous vigilance exercised by individuals.

In a number of domains, there are institutional procedures for evaluating the competence of individuals, making these evaluations public through some form of certification, and sanctioning false claims to being so certified. Medical doctors,

professors, judges, surveyors, accountants, priests, and so on are generally believed to be experts in their field because they have shown strong evidence of their expertise to experts who are even more qualified. Of course, these procedures may be inadequate or corrupt, and the domain may itself be riddled with errors; but still, such procedures provide clear and easily accessible evidence of an individual's expertise.

In analysing how information is assessed, filtered, and, in the process, transformed on the population scale, it is just as important to study vigilance towards the content as to study vigilance towards the source. We suggested above that vigilance towards the content is typically exercised through debate and argument, and may give rise to a kind of spontaneous division of cognitive labour. This division of labour can itself be culturally organized and take various institutional forms. Examples include judicial institutions, where a number of rules and procedures are designed to establish the facts of the matter through examination of the evidence, questioning of witnesses, and debates between the parties, for instance. The institutional organisation of epistemic vigilance is nowhere more obvious than in the sciences, where observational or theoretical claims are critically assessed via social processes such as laboratory discussion, workshops, conferences, and peer review in journals. The reliability of a journal is itself assessed through rankings, and so on (Goldman, 1999).

Social mechanisms for vigilance towards the source and vigilance towards the content interact in many ways. In judicial proceedings, for instance, the reputation of the witness is scrutinised in order to strengthen or weaken her testimony. In the sciences, peer review is meant to be purely content-oriented, but is influenced all too often by the authors' prior reputation (although blind reviewing is supposed to suppress this influence), and the outcome of the reviewing process in turn affects the

authors' reputation. Certification of expertise, as in the granting of a PhD, generally involves multiple complex assessments from teachers and examiners, who engage in discussion with the candidate and among themselves; these assessments are compiled by educational institutions which eventually deliver a reputation label, 'PhD', for public consumption.

Here we can do no more than point to a few of the issues raised by social mechanisms for epistemic vigilance. Our main aim in doing so is to suggest that, to a significant extent, these social mechanisms are articulations of psychological mechanisms linked through extended chains of communication, and, in some cases through institutional patterning (Sperber, 1996). In these population scale articulations, psychological mechanisms combine with cognitive artefacts (e.g., measuring instruments), techniques (e.g., statistical tests of confidence), and procedures (e.g., for cross-examination) to yield distributed epistemic assessment systems (Heintz, 2006) which should be seen as a special kind of distributed cognitive system (Hutchins, 1996).

The way in which people rely on distributed assessment systems poses a new version of Reid's and Hume's problem of how to justify our trust in tesminony. This is particularly true in the case of the new assessment systems without which we would be unable to use the Web at all. Google is a salient case in point. Google is not only a search engine, but is also used as an epistemic assessment engine. It implicitly represents, in the form of a ranked list, the relative epistemic values of Web documents found in a search. The higher the rank of a document, the more likely it is to contain relevant and reliable information. One way of producing such a ranking involves calculating the number of links to a given Web document from other Web sites, and weighting this number according to the relative importance of these sites

(which is itself calculated based on the number of links to them from still other sites). The idea behind this process is that linking to a document is an implicit judgement of its worth. The process compiles these judgements into an accessible indication: its rank on a search results page.

Why, then, do people rely on a search engine such as Google even though they know little or nothing about how its results pages are produced? And how far are they justified in doing so? Note, first, that our reliance is not entirely blind: This cognitive technology, like any other technology, is adopted on the basis of its observed success. Moreover, our reliance is tentative: We are willing to look first at highly ranked pages and to assume that there are good reasons why they are so highly ranked. But don't we then exercise some fairly standard epistemic vigilance towards the information we are presented with? The work and ideas evoked in this article should be relevant to an empirical investigation of such novel issues.

9. Concluding remark

Our aim in this paper has been to give some substance to the claim that humans have a suite of cognitive mechanisms for epistemic vigilance. To this end, we have surveyed issues, research and theories in different domains of philosophy, linguistics, cognitive psychology and the social sciences. We do not expect our readers to have accepted all our assumptions, several of which we ourselves view as rather speculative. What we do hope is to have a made a good case for the recognition of epistemic vigilance as an important aspect of human interaction. Just like communication, to which it is essentially linked, epistemic vigilance relies on individual mental mechanisms which

are articulated across individuals and populations into social mechanisms. Some of these mechanisms are targeted at the source of information, others at its content. Seeing these diverse mechanisms as all contributing to one and the same function of epistemic vigilance may be a source of insight in the study of each one of them.

Dan Sperber

Institut Jean Nicod

ENS, EHESS, CNRS, Paris

and

Department of Philosophy

Central European University

Budapest, Hungary

Fabrice Clément

Laboratoire de Sociologie

Département des Sciences Sociales

Université de Lausanne

Switzerland

Christophe Heintz

Department of Philosophy

Central European University

Budapest, Hungary

Olivier Mascaro

Cognitive Development Center

Department of Philosophy

Central European University

Budapest, Hungary

Hugo Mercier

PPE Program

University of Pennsylvania

Philadelphia, USA

Gloria Origgi

Institut Jean Nicod

ENS, EHESS, CNRS, Paris

Deirdre Wilson

Department of Linguistics

University College London

and

CSMN

University of Oslo

Norway

References

Adler, J. 2002: *Belief's Own Ethics*. Cambridge, MA: MIT Press.

Aikhenvald, A. Y. 2004: *Evidentiality*. Oxford: Oxford University Press.

Ainslie, G. 2001: *Breakdown of Will*. Cambridge: Cambridge University Press.

Alexander, R. D. 1987: *The Biology of Moral Systems*. Berlin: Aldine de Gruyter.

Asch, S. E. 1956: Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs*, 70, 1–70.

Austin, J. L. 1962: *How to Do Things withWords*. Cambridge, MA: Harvard UniversityPress.

Balci, F., Freestone, D. and Gallistel, C. R. 2009: Risk assessment in man and mouse. *PNAS*, 106, 2459–63.

Baron-Cohen, S., Leslie, A. M. and Frith, U. 1985: Does the autistic child have a 'theory of mind'? *Cognition*, 21, 37–46.

Bergstrom, B. and Boyer, P. Submitted: Who mental systems believe: effects of source on judgments of truth.

Bergstrom, B., Moehlmann, B. and Boyer, P. 2006: Extending the testimony problem: Evaluating the truth, scope, and source of cultural information. *Child Development*, 77, 531–8.

Bezuidenhout, A. 1998: Is verbal communication a purely preservative process? *The Philosophical Review*, 107, 261–288.

Birch, S., Vauthier, S. A. and Bloom, P. 2008: Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, 107, 1018–34.

Birch, S. A. J., Akmal, N. and Frampton, K. L. 2009: Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental Science*, 13, 363–69.

Blakemore, D. 1987: *Semantic Constraints on Relevance*. Oxford: Blackwell.

Blakemore, D. 2002: *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.

Bloch, M. 1998: *How We Think They Think: Anthropological Approaches to Cognition, Memory, and Literacy*. Boulder, CO: Westview Press.

Bond, C. F. Jr. and DePaulo, B. M. 2006: Accuracy of deception judgments. *Personality and Social Psychology Review*, 10, 214–34.

Boyer, P. 1992: *Tradition as Truth and Communication*. Cambridge: CambridgeUniversity Press.

Boyer, P. 2001: *Religion Explained: The Evolutionary Foundations of Religious Belief*. New York: Basic Books.

Boyd, R. and Richerson, P. J. 1985: *Culture and the Evolutionary Process*. Chicago, IL: Chicago University Press.

Burge, T. 1993: Content preservation. *Philosophical Review*, 101, 457–88.

Call, J. and Tomasello, M. 1999: A nonverbal false belief task: the performance of children and great apes. *Child Development*, 70, 381–95.

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

Carlson, S. M. and Moses, L. J. 2001: Individual differences in inhibitory control and children's theory of mind. *Child Development*, 72, 1032–53.

Carston, R. 2002: *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.

Chafe, W. and Nichols, J. eds. 1986: *Evidentiality: The Linguistic Coding of Epistemology*. Norwood, NJ: Ablex.

Chaiken, S. 1980: Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39, 752–66.

Cheney, D. L. and Seyfarth, R. M. 1990: *How Monkeys See the World*. Chicago, IL: Chicago University Press.

Clément, F. In press: To trust or not to trust? Children's social epistemology.*Review of Philosophy and Psychology*.

Clément, F., Koenig, M. A. and Harris, P. 2004: The ontogeny of trust. *Mind & Language*, 19, 360–79.

Van Cleve, J. 2006: Reid on the Credit of Human Testimony. In J. Lackey and E. Sosa (eds), *The Epistemology of Testimony*. Oxford: Oxford University Press.

Coady, C. A. J. 1992: *Testimony*. Oxford: Oxford University Press.

Corriveau, K. and Harris, P. L. 2009: Preschoolers continue to trust a more accurate informant 1 week after exposure to accuracy information. *Developmental Science*, 12, 188–93.

Corriveau, K. H. and Harris, P. L. In press: Young children's trust in what other people say. In K. Rotenberg (ed.), *Interpersonal Trust During Childhood and Adolescence*. Cambridge: Cambridge University Press.

Cosmides, L. and Tooby, J. 2005: Neurocognitive adaptations designed for social exchange. In D. M. Buss (ed.), *Evolutionary Psychology Handbook*. New York: Wiley, 584–627.

Couillard, N. L. and Woodward, A. L. 1999: Children's comprehension of deceptive points. *British Journal of Developmental Psychology*, 17, 515–21.

Csibra, G. and Gergely, G. 2009: Natural pedagogy. *Trends in Cognitive Sciences*, 13, 148–53.

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

Davidson, D. 1986: A coherence theory of truth and knowledge. In E. LePore (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell: 307–19.

Dawkins, R. and Krebs, J. R. 1978: Animal signals: information or manipulation? In J. R. Krebs and N. B. Davies (eds), *Behavioural Ecology: An Evolutionary Approach*. Oxford: Basil Blackwell Scientific Publications.

Denes-Raj, V. and Epstein, S. 1994: Conflict between intuitive and rational processing: when people behave against their better judgment. *Journal of Personality and Social Psychology*, 66, 819–29.

DePaulo, B. M. and Pfeifer, R. L. 1986: On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology*, 16, 249–67.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F. and Van Baaren, R. B. 2006: On making the right choice: the deliberation-without-attention effect. *Science*, 311, 1005–7.

Ekman, P. 2001: *Telling Lies*. New York: Norton.

Ekman, P. and O'Sullivan, M. 1991: Who can catch a liar. *American Psychologist*, 46, 913–20.

Evans, J. S. B. T. 2002: Logic and human reasoning: an assessment of the deduction paradigm. *Psychological Bulletin*, 128, 978–96.

Evans, J. S. B. T. and Frankish, K. (eds) 2009: *In TwoMinds*. Oxford:OxfordUniversity Press.

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

Figueras-Costa, B. and Harris, P. 2001: Theory of mind development in deaf children: a nonverbal test of false-belief understanding. *Journal of Deaf Studies and Deaf Education*, 6, 92.

Foley, R. 1994: Egoism in epistemology, in F. Schmitt (ed.), *Socializing Epistemology*. Lanham, MD: Rowman and Littlefield, Inc.: 53–73.

Fricker, E. 1995: Critical notice: telling and trusting: reductionism and anti-reductionism in the epistemology of testimony. *Mind*, 104, 393–411.

Fricker, E. 2006: Testimony and epistemic autonomy. In J. Lackey and E. Sosa (eds), *The Epistemology of Testimony*. Oxford: Oxford University Press.

Fusaro, M. and Harris, P. L. 2008: Children assess informant reliability using bystanders' non-verbal cues. *Developmental Science*, 11, 771–7.

Gee, C. L. and Heyman, G. D. 2007: Children's evaluation of other people's self-descriptions. *Social Development*, 16, 800–18.

Gilbert, D. T., Krull, D. S. and Malone, P. S. 1990: Unbelieving the unbelievable: some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59, 601–13.

Gilbert, D. T. and Malone, P. S. 1995: The correspondence bias. *Psychological Bulletin*, 117, 21–38.

Gilbert, D. T., Tafarodi, R. W. and Malone, P. S. 1993: You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65, 221–33.

Goldman, A. 1999: *Knowledge in a Social World*. Oxford: Oxford University Press.

Gouzoules, H., Gouzoules, S. and Miller, K. 1996: Skeptical responding in rhesus monkeys (Macaca mulatta). *International Journal of Primatology*, 17, 549–68.

Grice, H. P. 1957: Meaning. *Philosophical Review*, 66, 377–88.

Grice, H.P. 1969: Utterer's meaning and intentions. *Philosophical Review*, 78, 147–77. [Reprinted in Grice, 1989, pp. 86–116].

Grice, H. P. 1975: Logic and conversation. In P. Cole and J. P. Morgan (eds), *Syntax and Semantics, Vol. 3: Speech Acts*. New York: Seminar Press.

Grice, H. P. 1989: *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hardwig, J. 1985: Epistemic dependence. *The Journal of Philosophy*, 82: 335–349.

Harman, G. 1986: *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.

Harris, P. and Núñez, M. 1996: Understanding of permission rules by preschool children. *Child Development*, 67, 1572–91.

Hasson, U., Simmons, J. P. and Todorov, A. 2005: Believe it or not: on the possibility of suspending belief. *Psychological Science*, 16, 566–71.

Heintz, C. 2006: Web search engines and distributed assessment systems, *Pragmatics & Cognition*, 14, 387–409.

Henrich, J. and Boyd, R. 1998: The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, 19, 215–41.

Heyman, G. D. 2008: Children's critical thinking when learning from others. *Current Directions in Psychological Science*, 17, 344–7.

Holton, R. 1994: Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, 72: 63–76.

Hutchins, E. 1980: *Culture and Inference: A Trobriand Case Study*. Cambridge, MA: Harvard University Press.

Hutchins, E. 1996: *Cognition in the Wild*. Cambridge, MA: MIT Press.

Ifantidou, E. 2001: *Evidentials and Relevance*. Amsterdam: John Benjamins.

Jaswal, V. K., Croft, A. C., Setia, A. R. and Cole, C. A. In Press: Young children have a specific, highly robust bias to trust testimony. *Psychological Science*.

Kahneman, D. 2003: A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, 58, 697–720.

Koenig, M. A. and Echols, C. H. 2003: Infants' understanding of false labeling events: the referential roles of words and the speakers who use them. *Cognition*, 87, 179–203.

Koenig, M. A. and Harris, P. L. 2007: The basis of epistemic trust: reliable testimony or reliable sources? *Episteme*, 4, 264–84.

Koriat, A., Lichtenstein, S. and Fischhoff, B. 1980: Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory and Cognition*, 6, 107–18.

Krebs, J. R. and Dawkins, R. 1984: Animal signals: mind-reading and manipulation? In J. R. Krebs and N. B. Davies (eds), *Behavioural Ecology: An Evolutionary Approach*, 2nd edn. Oxford: Basil Blackwell Scientific Publications.

Kunda, Z. 1990: The case for motivated reasoning. *Psychological Bulletin*, 108, 480–98.
Lampinen, J. M. and Smith, V. L. 1995: The incredible (and sometimes incredulous) child witness: child eyewitnesses' sensitivity to source credibility cues. *Journal of Applied Psychology*, 80, 621–7.

Lewis, D. K. 1969: *Conventions*. Cambridge, MA: Harvard University Press.

Lewis, M., Stranger, C. and Sullivan, M. W. 1989: Deception in 3-year-olds. *Developmental Psychology*, 25, 439–43.

Locke, J. 1690 [1975]: *An Essay Concerning Human Understanding*, P. Nidditch, ed. Oxford: Oxford University Press.

Lutz, D. J. and Keil, F. C. 2002: Early understanding of the division of cognitive labor. *Child Development*, 1073–84.

Malone, B. E. and DePaulo, B. M. 2001: Measuring sensitivity to deception. In J. A. Hall and F. Bernieri (eds), *Interpersonal Sensitivity: Theory, Measurement, and Application*. Hillsdale, NJ: Erlbaum: 103–124.

Mann, S., Vrij, A. and Bull, R. 2004: Detecting true lies: police officers' ability to detect suspects' lies. *Journal of Applied Psychology*, 89, 137–149.

Mascaro, O. and Sperber, D. 2009: The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, 112, 367–80.

Mathew, S. and Boyd R. 2009: When does optional participation allow the evolution of cooperation? *Proceedings of the Royal Society of London B*, 276(1659), 1167–1174.

Matsui, T., Rakoczy, H.,Miura, Y. and Tomasello, M. 2009: Understanding of speaker certainty and false-belief reasoning: a comparison of Japanese and German preschoolers.

Developmental Science, 12, 602–13. Mercier, H. submitted-a: Developmental evidence for the argumentative theory of

reasoning.

Mercier, H. submitted-b: Looking for arguments.

Mercier, H. submitted-c: On the universality of argumentative reasoning.

Mercier, H. and Landemore, H. submitted: Reasoning is for arguing: consequences for deliberative democracy.

Mercier, H. and Sperber, D. 2009: Intuitive and reflective inferences. In J. St B. T. Evans and K. Frankish (eds), *In Two Minds*. New York: Oxford University Press.

Mercier, H. and Sperber, D. Forthcoming: Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Science*.

Milinski, M., Semmann, D. and Krambeck, H. J. 2002: Reputation helps solve the 'tragedy of the commons'. *Nature*, 415, 424–6.

Millikan, R. G. 1987: *Language, Thought and Other Categories*. Cambridge, MA: MIT Press.

Mills, C. M. and Keil, F. C. 2005: The Development of Cynicism. *Psychological Science*, 16, 385–90.

Morris, C.W. 1999: What is this thing called 'reputation'? *Business Ethics Quarterly*, 9, 87–102

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

Nowak, M. A. and Sigmund, K. 1998: Evolution of indirect reciprocity by image scoring. *Nature*, 383, 573–7.

Nurmsoo, E. and Robinson, E. J. 2009: Children's trust in previously inaccurate informants who were well or poorly informed: when past errors can be excused. *Child Development*, 80, 23–27.

Nurmsoo, E.,Robinson, E. J. and Butterfill, S. A. In press: Are children gullible? *Review of Psychology and Philosophy*.

Onishi, K. H. and Baillargeon, R. 2005: Do 15-month-old infants understand false beliefs? *Science*, 308, 255–58.

Origgi, G. 2005: A stance of trust. Paper presented at the *9th International Pragmatics Conference (IPRA)*, Riva del Garda, July 10–15th; to be published in T. Matsui (ed.):

Pragmatics and Theory of Mind. Amsterdam: John Benjamins (forthcoming). Origgi, G. 2008: Trust, authority and epistemic responsibility. *Theoria*, 23: 35–44.

Origgi, G. and Sperber, D. 2000: Evolution, communication and the proper function of language. In P. Carruthers and A. Chamberlain (eds), *Evolution and the Human Mind: Modularity, Language and Meta-Cognition*. Cambridge: Cambridge University Press.

Paglieri, F. and Woods, J. In press: Enthymematic parsimony. *Synthese*.

Pea, R. D. 1982: Origins of verbal logic: spontaneous denials by two-and three-year-olds. *Journal of Child Language*, 9, 597–626.

Perner, J. and Lang, B. 2000: Theory of mind and executive function: Is there a developmental relationship? In S. Baron-Cohen, H. Tager-Flusberg and D. Cohen (eds), *Understanding Other Minds: Perspectives From Autism and Developmental Cognitive*

Neuroscience, 2nd edn. Oxford: Oxford University Press.

Peskin, J. 1996: Guise and guile: children's understanding of narratives in which the purpose of pretense is deception. *Child Development*, 1735–51.

Petty, R. E. and Cacioppo, J. T. 1986: The elaboration likelihood model of persuasion. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*. Orlando, FL: Academic Press.

Povinelli, D. J. and Deblois, S. 1992: Young children's (Homo sapiens) understanding of knowledge formation in themselves and others. *Journal of Comparative Psychology*, 106, 228–38.

Reid, T. 1764 [2000]: *Inquiry into the Human Mind*, T. Duggan, ed. Chicago, IL: University of Chicago Press.

Richter, T., Schroeder, S. and W¨ohrmann, 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, 538–58.

Robinson, E. J., Haigh, S. N. and Nurmsoo, E. 2008: Children's working understanding of knowledge sources: confidence in knowledge gained from testimony. *Cognitive Development*, 23, 105–18.

Robinson, E. J. and Whitcombe, E. L. 2003: Children's suggestibility in relation to their understanding about sources of knowledge. *Child Development*, 74, 48–62.

Rosen, L. 1989: *The Anthropology of Justice: Law as Culture in Islamic Society*. Cambridge: Cambridge University Press.

Ross, L. 1977: The intuitive psychologist and his shortcomings: distortions in the attribution process. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*. *New York*: Academic Press.

Ross, L. and Nisbett, R. E. 1991: *The Person and the Situation: Perspectives of Social Psychology*. New York: McGraw-Hill.

Sabbagh, M. A. and Baldwin, D. A. 2001: Learning words from knowledgeable versus ignorant speakers: links between preschoolers' theory of mind and semantic development. *Child Development*, 72, 1054–70.

Schiffer, S. 1972: *Meaning*. *Oxford*: Clarendon Press.

Scofield, J. and Behrend, D. A. 2008: Learning words from reliable and unreliable speakers. *Cognitive Development*, 23, 278–90.

Searle, J. 1969: *Speech Acts: An Essay in the Philosophy of Language*, Cambridge: Cambridge University Press.

Shafir, E., Simonson, I. and Tversky, A. 1993: Reason-based choice. *Cognition*, 49, 11–36.

Shapin, S. 1994: *A Social History of Truth: Civility and Science in Seventeenth-Century England*, Chicago, IL: University of Chicago Press.

Southgate, V., Senju, A. and Csibra, G. 2007: Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18, 587–92.

Sperber, D. 1975: *Rethinking Symbolism*, Cambridge: Cambridge University Press.

Sperber, D. 1985: *On Anthropological Knowledge*: Cambridge University PressCambridge.

Sperber, D. 1996: *Explaining Culture: A Naturalistic Approach*, Oxford: Blackwell.

Sperber, D. 1997: Intuitive and reflective beliefs. *Mind & Language*, 12, 67–83.

Sperber, D. 2001: An evolutionary perspective on testimony and argumentation.

Philosophical Topics, 29, 401–13.

Sperber, D. In press: The Guru effect. *Review of Philosophy and Psychology*. Sperber, D. and Mercier, H. In press: Reasoning as a social competence. In J. Elster and H. Landemore (eds), *Collective Wisdom.*

Sperber, D. and Wilson, D. 1995: *Relevance: Communication and Cognition*, 2nd edn. Oxford: Blackwell.

Sperber, D. and Wilson, D. 2005: Pragmatics. In F. Jackson and M. Smith (eds), *Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press: 468–501.

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

Strawson, P. F. 1964: Intention and convention in speech acts. *The Philosophical Review*, 73, 439–60.

Sunstein, C. R. 2002: The law of group polarization. *Journal of Political Philosophy*, 10, 175–95.

Surian, L., Caldi, S. and Sperber, D. 2007: Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18, 580–86.

Talwar, V. and Lee, K. 2002: Development of lying to conceal a transgression: children's control of expressive behavior during verbal deception. *International Journal of Behavioral Development*, 26, 436–444.

Tesser, A. 1978: Self-generated attitude change. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*. New York: Academic Press.

Thagard, P. 2002: *Coherence in Thought and Action*, Cambridge, MA:MIT Press.

Tirole, J. 1996: A theory of collective reputations. *The Review of Economic Studies*, 63, 1–22.

Trommershauser, J., Maloney, L. T. and Landy, M. S. 2008: Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, 12, 291–97.

Vrij, A. 2000: *Detecting Lies and Deceit: The Psychology of Lying and. Implications for Professional Practice*. Chichester: John Wiley &. Sons.

Vrij, A. 2004: Why professionals fail to catch liars and how they can improve. *Legal andCriminological Psychology*, 9, 159–83.

Walker, M. B. and Andrade, M. G. 1996: Conformity in the Asch task as a function of age. *Journal of Social Psychology*, 136, 367–72.

Welch-Ross, M. K. 1999: Interviewer knowledge and preschoolers' reasoning about knowledge states moderate suggestibility. *Cognitive Development*, 14, 423–42.

Willingham, D. T. 2008: Critical thinking: why is it so hard to teach? *Arts Education Policy Review*, 109, 21–32.

Willis, J. and Todorov, A. 2006: First impressions: making up your mind after a 100-ms exposure to a face. *Psychological Science*, 17, 592–98.

Wilson, D. and Sperber, D. 2004: Relevance Theory. In L. Horn and G. Ward (eds), *The Handbook of Pragmatics*. Oxford: Blackwell.

Wimmer, H. and Perner, J. 1983: Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 41–68.

Yamagishi, T. 2001: Trust as a form of social intelligence. In K. Cook (ed.), *Trust in Society*. New York: Russell Sage Foundation.

Ybarra, O., Chan, E. and Park, D. 2001: Young and old adults' concerns about morality and competence. *Motivation and Emotion*, 25, 85–100.