Rev.Phil.Psych. (2012) 3:481–514
DOI 10.1007/s13164-012-0099-y

(Social) Metacognition and (Self-)Trust

Kourken Michaelian

Published online: 31 July 2012

Abstract What entitles you to rely on information received from others? What entitles you to rely on information retrieved from your own memory? Intuitively, you are entitled simply to trust yourself, while you should monitor others for signs of untrustworthiness. This article makes a case for inverting the intuitive view, arguing that metacognitive monitoring of oneself is fundamental to the reliability of memory, while monitoring of others does not play a significant role in ensuring the reliability of testimony.

1 Introduction

Sperber et al. (2010) have recently argued that humans are "epistemically vigilant" with respect to information communicated by other agents—roughly, that, given the evolutionary stability of communication, and given that it is often in the communicator's interest to deceive the receiver, receivers must have a capacity to filter out dishonest communicated information. This descriptive account dovetails, as I argue below, with the normative account defended by Fricker (1994, 1995) in the epistemology of testimony, according to which, while an agent is epistemically entitled to trust herself, she is not so entitled simply to trust the testimony of other agents but is required to monitor testifiers for dishonesty. Drawing on empirical and evolutionary work on communication and memory, this article argues that this "vigilantist" line

K. Michaelian (✉)
Felsefe Bölümü, Bilkent Üniversitesi,
FA Binası, Ankara 06800, Turkey
e-mail: kmichaelian@bilkent.edu.tr

gets things backwards: while agents neither need nor have a capacity for effective vigilance with respect to communicated information, they both need and have such a capacity with respect to internally generated information.

Section 2 reviews the theoretical framework employed throughout the paper. Section 3 briefly reviews vigilantism about testimony, and Section 4 argues that it should be rejected. Section 5 then develops a case for a form of vigilantism about memory and related internal sources.

2 The Metacognitive Epistemology Framework

My argument against vigilantism relies on a metacognitive epistemology framework (MEF) developed elsewhere (Michaelian 2012c, 2011c); this section briefly reviews the framework, showing how it can be extended to testimony via the concept of social metacognition.

2.1 Explaining Reliability

MEF foregrounds the explanatory question of how agents achieve epistemically acceptable belief-formation despite their dependence on imperfect information sources, focussing on the role of metacognitive monitoring in compensating for the limitations of such sources.

My focus here is on memory and testimony, and hence on metamemory and deception detection (which I treat as a form of social metacognition, playing a role analogous to that played by metamemory). The apparent analogy between memory and testimony make it useful to discuss them together. Typically, both are seen as purely “transmissive” sources, capable of preserving but not generating new justification/knowledge (Burge 1993; Senor 2007; Bernecker 2008). While this is not, strictly speaking, right (Lackey 2008; Matthen 2010; Michaelian 2011a, 2012b), the apparent analogy nevertheless provides a useful starting point, due in part to the way it breaks down. While reductionism, according to which the agent is not entitled simply to trust the deliverances of the relevant source but rather requires positive reason for doing so, is intuitively plausible in the case of testimony (Fricker 1994, 1995; Michaelian 2008) (though of course there are many defences of antireductionism about testimony in the literature, e.g., Burge 1993; Coady 1992), a form of antireductionism, according to which the agent is entitled to trust the deliverances of the source without requiring positive reason to do so, is intuitively plausible in the case of memory (Bernecker 2010). My argument here supports an inversion of this default position: I claim that, while the testimony/memory analogy does break down, it does so in an unexpected way, with agents requiring positive reason to trust memory but not testimony.

I set other internal sources, including perception and reasoning, aside. While research on metamemory (Dunlosky and Bjork 2008) and deception detection (Vrij 2008) is advanced, there is relatively little work on metaperception (Levin 2002; Loussouarn 2010), and consequently our understanding of metaperception is simply too limited at this point to enable us to discuss its epistemic role with much confidence. And while there is much work on metareasoning (Anderson et al. 2006; Cox 2005; Thompson 2010), metareasoning appears to play a role rather different from that played by other forms of metacognition: metareasoning seems to be primarily about allocating resources, choosing strategies, determining when to terminate a given strategy, etc. rather than about determining whether to accept/reject the product of a given cognitive process (though fluency does play a role here Oppenheimer 2008)—that is, about self-probing, rather than post-evaluation (Proust 2008). MEF would thus have to be extended significantly in order to take the epistemic role of metareasoning into account. This is anyway necessary, given that metamemory can also play a self-probing role (e.g., the feeling of knowing can determine whether the agents continues to attempt to retrieve a record Koriat 1998; de Sousa 2008; Dokic 2013), but it is a separate project.

The normative component of MEF incorporates process reliabilism about justification (Goldman 1979): the degree of justification of a belief is determined by the reliability of the process that produces it, where the reliability of a process is defined as its tendency to produce a given ratio of true beliefs to total (true and false) beliefs. I assume that epistemically acceptable belief-formation requires a high level of reliability. MEF also makes room for other epistemic desiderata, especially power and speed (drawing inspiration from Goldman 1992; Cummins et al. 2004; Lepock 2007), incorporating these into a broadly virtue-reliabilist framework (Sosa 2007). The core idea is that an appropriate balance of reliability, power, and speed, where the appropriate balance is determined by the function of the relevant cognitive system together with the agent's current context, including such factors as the relative importance of forming true beliefs and avoiding false beliefs, is required for virtuous belief-formation. I discuss power and speed briefly below, but my focus here is largely on reliability.

MEF focusses on belief-producing processes with a two-level structure, in which an *information source* produces representations that, by default, come to serve as belief-contents and an *endorsement mechanism* determines whether to endorse or reject produced representations—the endorsement mechanism in effect functions as a filter on the information source. It is useful to conceive of the endorsement mechanism as employing a set of criteria for evaluating produced representations, together with a rule determining whether a given representation is to be evaluated as accurate given the extent to which it satisfies these criteria; this rule can be flexible, requiring greater or lesser satisfaction of the criteria according to the agent's current context, thus allowing trade-offs among reliability, power, and speed.

The way in which the reliability R of a two-level process is determined can be understood using a modified version of a probability model originally developed by Park and Levine to understand what determines deception detection accuracy (Park and Levine 2001). Assuming that a representation is endorsed only if it is evaluated as accurate, R is determined as follows, where T means that the information source produces an accurate representation and H

means that the endorsement mechanism evaluates the representation as being accurate:

$$R = P(H\&T)/P(H\&T) + P(H\& \sim T) \tag{1}$$

This is just the ratio of accurate evaluations of representations as accurate to total (accurate and inaccurate) evaluations of representations as accurate. $P(H\&T)$ and $P(H\& \sim T)$ are determined by the base rate $P(T)$ of accurate representations produced by the source and the sensitivity of the endorsement mechanism to accuracy and inaccuracy, $P(H|T)$ and $P(\sim H| \sim T)$.

Figure 1 illustrates two key points that can be made using this probability model: (1) if the endorsement mechanism is sufficiently sensitive (high $P(H|T)$ and $P(\sim H| \sim T)$), a reliable information source is unnecessary—effective monitoring renders reliable content-production unnecessary for reliable belief-production; (2) if the base rate of accurate representations is sufficiently high (high $P(T)$), a sensitive endorsement mechanism is unnecessary—reliable content-production renders effective monitoring unnecessary for reliable belief-production. Both points are straightforward but worth making explicitly, since they have typically been neglected in the epistemologies of memory and testimony.

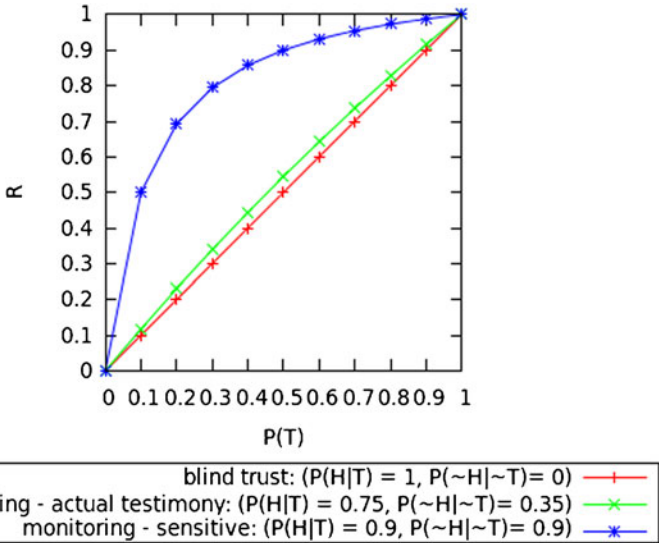

Fig. 1 Illustration of how the reliability of a two-level belief-producing process is determined by the interaction between the base rate of accurate representations produced by its information source and the sensitivity of its endorsement mechanism to the accuracy/inaccuracy of those representations. A highly sensitive endorsement mechanism can compensate for a low base rate, whereas a high base rate can compensate for an insensitive endorsement mechanism. If the agent trusts the information source blindly, accepting all incoming representations, reliability is simply determined by the base rate. In the case of testimony, our actual monitoring is barely sensitive, so that a high base rate of honest testimony is necessary for reliable belief-production (see Section 4)

2.2 The Role of Metacognition

In general, the role of the endorsement mechanism should be understood in terms of metacognition. (Metacognition appears to be evolutionarily recent, restricted to humans and perhaps a few other species Metcalfe 2008; Smith et al. 2003; Proust 2006, so we should expect that non-metacognitive animals will not in general have two-level systems.) Classically understood (Nelson and Narens 1994), metacognition refers to the monitoring and control of mental processes: a meta-level monitors an object-level, updating its model of the latter, on the basis of which it controls the activity of the object-level. Since endorsement is a matter of control, a two-level system is metacognitive if endorsement decisions are based on information received by the endorsement mechanism about the operation of the information producer.

We can distinguish among three types of information about the object-level to which an endorsement mechanism can have access: (1) general knowledge about the source, e.g., about its overall reliability; (2) information about the operation of the source during the production of a given representation, e.g., the speed with which the representation was produced; (3) information about a produced representation itself, e.g., about features of its content or its relation to other representations. We can refer to metacognition based on the latter two types of information as cue-based and to metacognition based on the first type as knowledge-based.[1] Cue-based metacognition is typically a type 1 (automatic, heuristic, unconscious, fast) process, while knowledge-based metacognition is typically a type 2 (systematic, reflective, conscious, slow) process (Evans 2008; Frankish 2010).

The basic structure of a metacognitive belief-producing process is summarised in Fig. 2.

MEF treats testimony as an epistemic source on a par with internal sources, with monitoring for cues to deception viewed as a form of social metacognition.[2] I discuss monitoring for cues in Section 4; here, I discuss the more controversial move of treating receipt of testimonial information as analogous to receipt of representations produced by internal sources, as a process which, by default, leads to endorsement and hence production of a belief having the relevant representation as its content. (Default endorsement is compatible with vigilance—the claim that endorsement is the default means that agents tend to endorse received information if monitoring, assuming that it is engaged, does not produce reason to reject it.)

[1] The cue-based/knowledge-based distinction is not quite the same as Koriat's distinction between experience-based and theory-based metacognition (Koriat 2006), since there is no requirement that sensitivity to cues manifest itself in the form of a feeling.

[2] Jost et al. (1998) also refer to social metacognition, but they justify this by employing an extremely broad definition of metacognition as involving "any aspect of thinking about thinking", a definition so broad as to include many entirely disparate phenomena; my conception of social metacognition is much narrower.

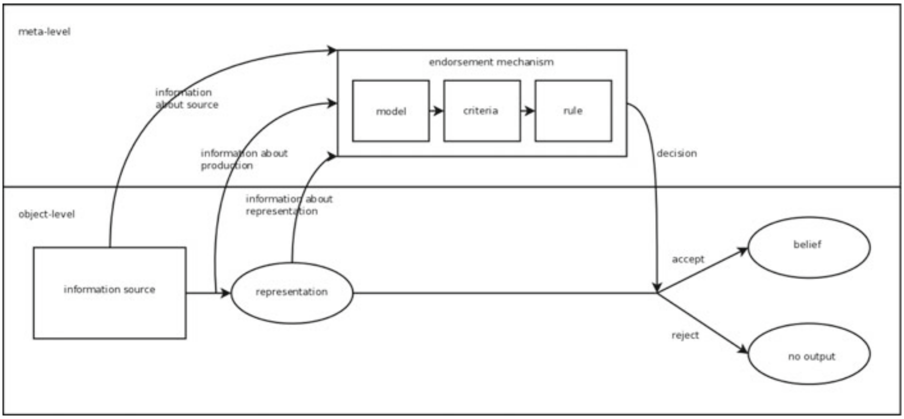

Fig. 2 Basic structure of a generic metacognitive belief producing process

Reid famously suggested that we have “a disposition to confide in the veracity of others, and to believe what they tell us” (Reid 1764, p. 238–40). This suggestion would license a picture on which the receipt of testimony tends by default to eventuate in the production of the corresponding belief. While Reid’s suggestion is speculative, the work of Gilbert and his colleagues (Gilbert et al. 1990, 1993; Gilbert 1991) provides some support for it. Gilbert argues that belief-formation is not a separate step following comprehension but rather inevitably accompanies comprehension, with rejection being an optional later step:

> Findings from a multitude of research literatures converge on a single point: People are credulous creatures who find it very easy to believe and very difficult to doubt. In fact, believing is so easy, and perhaps so inevitable, that it may be more like involuntary comprehension than it is like rational assessment. (Gilbert 1991, p. 117).

Since Sperber et al. (2010) challenge the evidence for Gilbert’s view, citing studies (Hasson et al. 2005; Richter et al. 2009) that suggest that the tendency for automatic acceptance is overcome when the communicated information is relevant to the recipient, however, it risks being question-begging to rely on Gilbert in this context.

While I will not respond respond to the details of the Sperber et al. critique of Gilbert’s view here, I take it to be shown by work by deception researchers on the truth bias, defined as the tendency of recipients to evaluate messages as honest regardless of actual message honesty (Park and Levine 2001; McCornack and Parks 1986; Levine et al. 1999, 2006; Bond and DePaulo 2008; Levine and Kim 2010), that a picture along roughly Gilbertian/Reidian lines must be right, whether or not the details of Gilbert’s view are right. This work shows clearly that, while “truth-bias varies in degree from person to person and situation to situation”, “most people are truth-biased most of the time" (Levine and Kim 2010). Levine and Kim cite several factors which might together explain the truth bias, including belief as a mental default (basically Gilbert's picture), the necessity of truth bias for communication (drawing on Grice 1989), and the necessity of truth bias for social interaction; I offer an alternative explanation, in terms of the costs of deception, in Section 4.3.2. Whatever the details of the mechanism(s) giving rise to the truth bias, however, the existence of the bias is enough to justify treating endorsement of testimonial information as the default, with monitoring for cues to deception then playing a filtering role analogous to that played by metamemory with respect to information retrieved from memory.

This allows MEF to treat belief-formation in memory and testimony as analogous, as far as the structure of the relevant processes is concerned. *Memory:* Memory retrieval is the information source, with retrieved (apparent) memories endorsed by default. The agent might monitor his memory, the production of the relevant memory representation, or the representation itself. Monitoring is typically a cue-based, type 1 process. The reliability of memory belief-formation is determined by the interaction between (1) the base rate of accurate apparent memories and (2) the sensitivity of the agent to the accuracy/inaccuracy of apparent memories. *Testimony:* The communicator is the information source. The receiver monitors the communicator, the production of the relevant utterance, or the utterance itself. Setting aside monitoring of the speaker himself (knowledge-based metacognition), which plays a relatively minor role (see Section 4), what matters is cue-based monitoring of the production of the utterance or of the utterance itself; this will normally be accomplished by type 1 processes. The reliability of testimonial belief-formation is determined by the interaction of (1) the base rate of accurate testimony and (2) the sensitivity of the receiver to the accuracy/inaccuracy of received testimony.

3 Vigilantism About Testimony

This section provides an overview of vigilantism about testimony; Section 4 then applies MEF to argue against vigilantism.

I do not mean to assimilate Sperber et al.'s view on epistemic vigilance to Fricker's brand of local reductionism. Fricker and Sperber can be grouped together under the heading of vigilantism, despite the fact that they focus on somewhat different questions, because their positions are representative of a certain broad view of the epistemic role of communicated information. On this view, testimony is to be sharply distinguished from internal epistemic sources: it is, in some sense, less reliable than internal sources and thus requires a level of surveillance not required by internal sources. This view has a good deal of intuitive plausibility, for, as Sperber points out, while other agents regularly have an incentive to deceive us, presumably our own internal information sources do not.

Fricker focusses on a pair of related epistemological questions: Is it necessary to monitor in order to form justified testimonial beliefs? Is it possible to monitor in a way that permits formation of justified testimonial beliefs? While these questions are normative, they are, to the extent that we take justification to be a matter of reliability, at the same time descriptive, and they are thus very close to Sperber's descriptive question about the role of monitoring in the formation of testimonial beliefs. While Sperber et al. are not interested primarily in the role of monitoring in generating epistemic justification, they are interested in the role of monitoring in determining the reliability of testimonial belief-formation, and this is enough to license grouping the two positions together under the heading of vigilantism—ultimately, similar claims about the capacity of monitoring to filter out deceptive testimony are at the heart of the two positions.

Fricker's and Sperber's arguments focus on what Fricker refers to as trustworthiness, where an agent's testimony is trustworthy on a given occasion when she is both honest and competent with respect to the topic of her testimony on that occasion. Sperber assumes that agents are largely competent: "individual mechanisms are, under normal conditions and in the absence of social interference, reliable sources of true beliefs" (Sperber 2001) (see also Sperber et al. 2010, p. 359). I will likewise assume that agents' individual belief-forming processes are reliable, an assumption licensed by basic evolutionary considerations (McKay and Dennett 2009), therefore focussing on honesty rather than competence. Similarly, with Sperber, I assume that testimonial beliefs, like other beliefs, must in general be true to be advantageous: just as we can set aside non-testimonial cases in which biased belief-formation is beneficial (e.g., self-enhancement Alicke and Sedikides 2009) as being marginal, we can set aside cases in which acceptance of inaccurate testimony is beneficial (e.g., exaggerated warnings Sperber 2001).

Both Sperber and Fricker assign a role to type 2 deception detection: recipients will sometimes have antecedent reason not to trust the relevant speaker (on the relevant occasion, on the relevant topic), and they will sometimes later discover that the speaker lied. But, in practice, such knowledge-based deception detection appears to play a relatively minor role. First, it seems likely that lies are relatively seldom detected be means of antecedent knowledge, for, in most cases, we simply lack the necessary antecedent knowledge: we often receive testimony from speakers about whom we have little or no information; moreover, even when we have some information about the speaker, this need not tell us whether he is likely to be honest on a given occasion, since honesty is determined as much by situational factors as the agent's dispositions, so that past honesty provides a poor guide to future honesty. Second, it seems likely that lies are relatively seldom detected after the fact, when the speaker discovers the deception by other means (information received from others, physical evidence, etc.). Many lies will go undetected, since there is no relevant evidence. In certain cases, even when evidence is available, the subject might (for a variety of reasons) more or less actively avoid discovering a lie (Vrij 2008). Finally, poor source memory raises an addition problem for after-the-fact lie detection: even if the subject discovers after the fact that the source was dishonest, the belief resulting from acceptance of the testimony can persist (Kumkale and Albarracín 2004). I therefore set knowledge-based detection aside here, focussing on cue-based detection occurring at the time at which the relevant utterance is received.[3]

3.1 Local Reductionism

Fricker[4] contrasts her local reductionism with antireductionism, the thesis, roughly, that agents are entitled to assume trustworthiness (and so honesty), and therefore to form a testimonial belief, absent special reason not to. According to reductionism in general, the agent must have positive reason to hold that the testifier is trustworthy and therefore to form a belief by accepting his testimony; according to local reductionism in particular, the agent acquires such reason by monitoring the testifier for signs of untrustworthiness, including dishonesty. Thus, while antireductionism licenses what we can refer to as a policy of blind trust, local reductionism requires what we can refer to as a monitoring policy, a policy of refraining from the formation of testimonial belief unless one has monitored the speaker for signs of untrustworthiness, and unless no such signs have been found.

Against antireductionism, Fricker argues that a policy of blind trust is a recipe for gullibility, unreliability.[5] She therefore concludes that subjects must, if they are to form justified testimonial beliefs, use the monitoring policy. She argues, moreover, that we do in fact use the monitoring policy, and that the policy is effective, that is, that it actually enables us to screen out untrustworthy testimony, thus enabling reliable testimonial belief production in the face of unreliable testimony.

In order for these two arguments to work, Fricker requires several empirical assumptions about lying:

A1 Dishonest testimony is frequent, so that blindly accepting testimony would lead to unreliable testimonial belief-formation, due to acceptance of dishonest and therefore false testimony.

A2 Reliable cues to dishonesty exist.

[3] As Gelfert has argued (Gelfert 2009), reliance on type 2 monitoring may result in local reductionism collapsing into Reidian credulism, as the sort of innate, subpersonal mechanisms on which type 2 monitoring relies are the very mechanisms invoked by Reid to explain how testimonial knowledge is acquired. This objection is compatible with my argument here.

[4] I rely mainly on Fricker (1994, 1995), but see Fricker (2002, 2004, 2006a, b, c) for recent developments of the view. I have provided a detailed reconstruction of Fricker's epistemology elsewhere Michaelian (2010); I rely on that reconstruction here without defending it.

[5] While Fricker is an internalist, she acknowledges the importance of reliability, at least as I read her. I set aside the internalist (coherentist) aspect of her argument, as this is irrelevant given MEF. There is also a modal aspect to Fricker's argument—the blindly trusting subject is supposed to be gullible not only in the sense that her beliefs are formed by an unreliable process but also in the sense that they are unsafe or insensitive. I set this aspect of the argument aside here, as I have dealt with it elsewhere Michaelian (2010).

A3 Subjects can and do effectively monitor for (enough of) these cues, and not for (too many) false cues. That is, they "look" for appropriate cues, and they can detect them when they are present.

3.2 Epistemic Vigilance

In addition to their evolutionary argument, Sperber et al. cite a number of more direct sources of apparent evidence for their view, e.g., experimental work showing that children show a preference for benevolent informants (Mascaro and Sperber 2009). While this work is interesting, it is not directly relevant to the question whether we are vigilant by default: that children prefer to rely on testimony from informants who are already known to be benevolent suggests that we are prone to distrust a communicator when we have antecedent reason to expect her to be dishonest but not that we have any sensitivity to her dishonesty when we have no such antecedent information. I therefore focus on their evolutionary argument.[6]

We can distinguish two main readings of the argument. On the first, they argue as follows; since it is often in the interests of communicators to deceive, we should assume that deception is frequent; we should also assume that testimonial belief-formation is reliable, since, otherwise, it would not be evolutionarily stable; so we should conclude that receivers are epistemically vigilant, and, in particular, that they are able to filter out most dishonest testimony. They are not, however, very explicit on the supposed frequency of deception. Since it is more conservative not to assume that deception is frequent, and since, as we shall see, there is evidence that, while deception does occur regularly, deceptive utterances account for a small fraction of communication (Levine and Kim 2010; DePaulo et al. 1996; Levine et al. 2010; Serota et al. 2010), there is reason to prefer a reading of the argument which does not require the assumption that deception is frequent. On this version of the argument, while deception might be infrequent, nevertheless it does occur regularly; so agents who effectively monitor for deception will enjoy an evolutionary advantage; so we should expect to find that agents are epistemically vigilant with respect to deception.

The empirical presuppositions of Sperber's view are similar to those of Fricker's. The first version of the argument, on which we should expect agents to be epistemically vigilant with respect to deception because it is often in the interests of communicators to lie, works only if communicators do in fact lie frequently, so that blind acceptance of testimony would be unreliable—if the interest of communicators in lying does not translate into frequent lying, then blind acceptance of testimony will be reliable, and there will be no need to screen out dishonest testimony. The argument likewise requires A2 and A3—otherwise, vigilance will not be effective.

[6] I rely here on the more detailed critique given in Michaelian (2012a).

The second version of the argument does not require that deception is frequent enough to render testimonial belief-formation by blind trust unreliable but only:

A1* Deception occurs regularly.

Only this weak assumption is required for it to be the case that agents who monitor for deception enjoy an evolutionary advantage. Nor does it require that cues enabling highly effective monitoring exist or that agents exploit these to monitor effectively, though it does require:

A2* Cues enabling monitoring good enough to produce some benefit exist.
A3* Subjects can and do monitor for these cues, reliably enough to produce some benefit.

But because this is an argument about the adaptivity of different possible mechanisms for testimonial belief-formation, it requires an additional assumption, namely:

A4 The benefits of filtering out deception (when it does occur) outweigh the costs of monitoring for deception.

4 Against Vigilantism About Testimony

In this section, I argue that, when we assess these assumptions against evidence from research on human communication, it becomes clear that vigilantism about testimony is untenable. I begin by focussing on the assumptions required by Fricker's argument and the first reading of Sperber's argument; I then discuss the assumptions required by the second reading of Sperber's argument.

4.1 What is the Base Rate of Deception?

A1 says that there is a high base rate of dishonest testimony. The problem here is straightforward: there are both empirical and theoretical reasons to expect that the base rate of dishonest testimony is low.

4.1.1 Empirical Evidence

While it is difficult to see how the base rate of deception might be empirically determined with much accuracy (this parallels the difficulty of determining the base-rate of accurate representations produced by memory retrieval—see Section 5.3), existing empirical attempts to get some sense of the frequency of lying in non-laboratory settings (the base rate is usually set artificially to 50 % in deception detection experiments) all suggest that lying is a regular but infrequent occurrence, with most people lying infrequently (Levine and Kim 2010; DePaulo et al. 1996; Levine et al. 2010; Serota et al. 2010).

4.1.2 Theoretical Considerations

While tentative, these findings fit well with what we should expect on evolutionary grounds. In the case of animal signalling, the following reasoning is more or less standard (Searcy and Nowicki 2005): in order for signalling to be evolutionarily stable, it must be beneficial not only to senders but also to receivers, for, if signalling were not beneficial to receivers, they would stop accepting signals, and consequently signals would stop being given; this requires that a majority of signals are honest (since usually a signal has to be honest in order to be useful to the recipient). Discussion then focusses on the means by which the honesty of signalling is ensured.

One well-known explanation of the honesty of signalling is provided by the handicap principle, according to which signals are hard to fake because costly, which guarantees that they will usually be honest, since dishonest signallers will bear a burden that they cannot support (Grafen 1990; Zahavi and Zahavi 1997). But there are other mechanisms which can also ensure that signals are mostly honest: a signal can be an index, i.e., its meaning can be tied to its form; and signalling can be kept honest through deterrents, in which case costs are paid by dishonest signallers (as opposed to honest signallers, as in the handicap principle) (Scott-Phillips 2008).

Caution is required when attempting to apply this standard reasoning to human communication, for, as Sperber points out (Sperber 2001), there are some important differences between the two cases. First, there is the point that human communication is typically cheap (unlike, say, the peacock's tail): it does not cost much to produce a typical testimonial utterance, so the handicap principle does not get a grip. That the handicap principle cannot account for the reliability of communication does not, however, mean that communication need not be reliable in order to be evolutionarily stable or that its reliability cannot be explained by the operation of another mechanism. And while the handicap principle does not get a grip here, deterrence does—I turn to this point next, exploring the costs of lying (relative to honesty); evidence that lying is significantly more costly than honesty comes from a variety of sources.

Second, humans have the capacity for theory of mind, which allows them to engage both in more sophisticated attempts to deceive and in more sophisticated attempts to counter deception. In particular, the mindreading capacity of receivers can in principle increase their capacity to monitor effectively, and this might seem to require a change to the basic evolutionary reasoning—if effective monitoring is available to recipients, they can filter out dishonest testimony, and so it need not be the case that testimony is honest on average. This overlooks the point that detection of deception will often result in punishment of the deceiver, which functions as a deterrent to deception. More importantly, it overlooks the point (discussed in Section 4.3 below) that our mindreading capacity does not as a matter of fact enable us reliably to detect deception. Thus it still needs to be the case that the base rate of deception is low, despite mindreading—otherwise, communication would not be advantageous to receivers, and it would collapse.

Of the available mechanisms, it appears most likely that human communication is kept honest through deterrents—in particular, that the cognitive, psychological, and social costs of lying mean that lying is, on average, significantly more costly than honesty.

Cognitive costs of deception Vrij et al. argue that, for a number of reasons (including the need to generate a coherent lie, the need to monitor the receiver, and the need to both activate the lie and suppress the truth), deception is cognitively more costly than honesty (Vrij et al. 2011, p. 28–29). This view receives support from the work of Vrij's group showing that lie detection accuracy is improved when cognitive load is increased (Vrij 2008; Vrij et al. 2011, 2006, 2008), as well as imagining studies showing no area of the brain more active for honesty than for deception (Christ 2009; Spence and Kaylor-Hughes 2008; Verschuere et al. 2011). A view of lying as cognitively costly also fits well with the view that self-deception evolved to facilitate interpersonal deception, since this claim is supported in part by the point that self-deception eliminates the cognitive load that would otherwise be involved in other-deception (von Hippel et al. 2011).

Psychological costs of deception In addition to the cognitive costs of lying, we should not overlook the point that lying will, in normal agents, have a "psychological" cost, since we internalize norms that forbid lying, except under special circumstances. These norms are widespread, apparently pan-cultural: (The Global Deception Research Team 2006) cites a world values survey (Inglehart et al. 1998) in which 48 % of respondents said that lying in one's self-interest is never justified; a newer version of the survey gives 46 % (Inglehart et al. 2004). Violation of internalized norms constitutes another disincentive to lying.

Social costs of deception The existence of norms against lying matters in another way, since it means that lies, if detected, will often be punished. There is thus a social cost to lying: though an agent might know in some cases that his lie will not be detected, he must also, when he is unsure of this, take into account the losses (damage to relationships, loss of reputation, etc.) that are likely to occur if a lie is detected.

Summing up: The direct empirical evidence that lying is infrequent, the need to explain the evolutionary stability of communication, and considerations suggesting that lying is more costly than truth-telling suggest that, contra A1, the base rate of deception is low. This means that there is likely no need for a monitoring mechanism to filter out dishonest testimony in order to achieve an acceptable level of reliability in testimonial belief-formation—the base rate $P(T)$ can do all the work. As far as the second version of Sperber's argument is concerned, it means that monitoring will have to be not only effective but also cheap in order to be adaptive (I develop this point below). The assumption that deception occurs regularly (A1*), however, is clearly safe.

4.2 Are There Reliable Cues to Deception?

A2 requires the existence of reliable cues to deception, cues that can actually be used to detect deception in real time. However, based on extensive review of studies, Vrij (2008) concludes that there are only a small number of cues to deception, and these are rather weak—the cue sometimes accompanies deception but not always, and sometimes is present when the communicator is honest. The same point holds with respect to nonverbal cues. Thus we can expect that it is difficult for agents to use these cues to detect deception. As Vrij points out, there are a number of possible explanations for the fact that genuine cues to deception are weak and few, including interpersonal differences (different subjects behave differently when lying—e.g., lying might be easier for more intelligent people) and situational factors (different situations induce different behaviours—e.g., by affecting motivation). But for present purposes, what matters is only the basic point that there are only a few weak cues to deception.[7]

It remains possible, despite this point, that agents manage to exploit the available cues to monitor effectively for deception. This might occur if they are sensitive to enough of the genuine cues (and not to too many false cues) and rely on an appropriate set of cues. So it is not entirely clear whether A2 is correct (I return to A2* below); as we will see, however, even if it is, A3 is not.

4.3 Is Monitoring for Deception Effective?

A3 says that agents use the available cues to monitor effectively for deception. Though there are only a few weak cues to deception, agents might nevertheless, through monitoring for a combination of cues, exploit the existing cues to monitor effectively. However, while folk psychology suggests that we are able to detect deception reasonably well, this intuition is not born out by empirical work on deception detection.[8]

4.3.1 Empirical Evidence

In an early review of deception detection studies, Kraut found an average accuracy rate of 57 % (Kraut 1980); more recently, Vrij estimated 56.6 % (Vrij 2000), while Bond and DePaulo estimated 54 % (Bond and DePaulo 2006); in the 2008 edition of his book, Vrij gives an accuracy rate of 54.25 % (Vrij 2008).

[7] Douven and Cuypers (2009) similarly point out that Fricker may overestimate the availability of cues to untrustworthiness.

[8] As noted in Section 4.4.1 below, what affects the reliability of testimonial belief formation is not the overall reliability of evaluations of testimony as honest or dishonest but rather the reliability specifically of evaluations of testimony as honest; but the finding of poor deception detection accuracy establishes that monitoring for deception is not effective, so I set this aside for now.

Here, I will follow Vrij in taking accuracy to be around 54 %,[9] but the precise number does not matter for my argument.

A natural worry about this finding is that it is limited to interactions with unfamiliar communicators: it is plausible that, when we have more background information about a communicator, we will better be able to detect his lies. But this is not born out by empirical work—there appears to be no real difference between deception detection accuracy with respect to strangers and accuracy with respect to spouses, family members, etc. (Vrij 2008; McCornack and Parks 1986; Levine et al. 1999; McCornack and Levine 1990; Millar and Millar 1995; Anderson et al. 2002). This might be because a communicator who knows the recipient is better able to craft a lie designed to fool that particular recipient (as Vrij suggests), but the precise explanation does not matter here—what matters is only that being acquainted with the communicator does not confer any advantage on the recipient, as far as ability to detect deception is concerned. Moreover, given the extent of our interactions with communicators about whom we have no prior information, the objection does not go very far in challenging the claim that the vigilantist assumption is empirically untenable.

4.3.2 Theoretical Considerations

In other words, deception detection accuracy is consistently found to be barely above chance—recipients do not succeed in effectively monitoring communicators for deception, contrary to A3, and thus they cannot filter out dishonest testimony. This extremely robust finding can be explained in part in terms of inaccurate folk psychological beliefs about deception and in part in terms of the operation of the truth bias.

If we assume that an agent's deception judgements correspond to her beliefs about cues (Vrij 2008; Forrest et al. 2004), then inaccurate beliefs about cues can go some way towards explaining poor deception detection accuracy. And indeed, work on beliefs about cues confirms that folk psychological beliefs about cues are largely inaccurate—there is little overlap between believed cues and genuine cues. For example, an ambitious study, involving subjects in 75 different countries, speaking 43 different languages, identified a cross-cultural stereotype of a liar, a stereotype which is largely inaccurate (The Global Deception Research Team 2006): e.g., by far the most commonly-reported belief about cues to deception is that liars avert their gaze; however, gaze aversion is not a cue to deception (DePaulo et al. 2003).

Summarizing existing work on beliefs about cues to deception, Vrij points out that, overall, though there is some overlap, there is little correspondence between beliefs about cues to deception and actual cues. Of twenty-four "cues" he considers, folk psychology is right about six. Of these six, however, only three can serve as genuine cues to deception; the rest bear no relationship

[9] The accuracy rate needs to be relativized to the base rate of 50 % honest statements; I come back to this below.

to deception, and so accurate beliefs here do not aid in deception detection. In contrast, folk psychology sees many cues where none exist: Vrij identifies eleven false cues to deception, behaviours that are unrelated to deception but which are believed to be related to deception. Moreover, folk psychology overlooks many genuine cues to deception, and in some cases assigns a meaning to a cue opposite to its real meaning (e.g., liars tend to make few hand and finger movements, but are believed to make more hand and finger movements). Given the weakness of the few genuine cues to deception, deception detection is a difficult task; given the inaccuracy of beliefs about cues, it is unsurprising that deception detection accuracy is poor.

In addition to inaccurate beliefs about cues, the role of the truth bias also needs to be taken into account in explaining poor deception detection accuracy, as Park and Levine and their collaborators have shown in a series of papers. In order to account for the veracity effect (in which detection accuracy is a function of message honesty) (Levine et al. 1999), Park and Levine posit that subjects are in general truth-biased (Park and Levine 2001; Levine et al. 2006; Levine and Kim 2010). Plugging this assumption into their probability model allows the model to predict deception detection accuracy rates well. Where H means that the agent judges that the received testimony is honest and T means that the received testimony is in fact honest, overall deception detection accuracy is given by $P(H\&T) + P(\sim H\& \sim T)$. Reflecting a strong truth bias, Park and Levine fix $P(H|T) = 0.779$ and $P(\sim H| \sim T) = 0.349$; if the base rate of honesty is 0.5, the model then predicts an accuracy rate of about 56 %. Thus the truth bias contributes to explaining the ineffectiveness of monitoring for deception.

This raises the question of how to account for the existence of the truth bias itself. Several factors seem to be at work here, including the prevalence of false beliefs about cues to deception (already discussed above); in addition to this, we should take into account the costs of monitoring for deception, which appear to roughly parallel the costs of lying, as well as the adaptivity of a disposition to judge received messages as honest.

Cognitive costs of monitoring Monitoring requires cognitive resources over and above those required for mere interpretation of an utterance, so there will in general be an incentive not to monitor, if possible. Together with the need to rely on information communicated by others, this can be expected to contribute to a tendency to evaluate received messages as honest.

Psychological costs of monitoring Just as there are psychological costs involved in lying, there are psychological costs associated with being on the lookout for deception, and for similar reasons: just as we internalize norms against lying, we internalize norms against excessive suspicion. Additionally, being on the lookout for deception might tend to heighten the agent's awareness of his vulnerability to manipulation through deception—given his inability to do much about this, it might be preferable for him to be less aware of the possibility, just as we tend to overestimate ourselves in many other domains.

Social costs of monitoring The norm against excessive suspicion means that there can be a social cost to monitoring—if one is seen to be too untrusting, this can result in social sanctions. Additionally (this is noted by Vrij, among others), the very rules of polite conversation tend to make monitoring more difficult—e.g., certain types of questioning can aid in uncovering deception, but these types of questioning are prohibited in normal conversations. Thus, subjects have an additional incentive to simply accept communicated information rather than monitor for dishonesty.

Adaptivity of the truth bias Given that the base rate of deception is low, it is plausible that a disposition to evaluate received testimony as honest has been selected for. As Park and Levine point out (Levine et al. 2006), given that lying is an infrequent occurrence, it is likely adaptive for agents to accept most received testimony, remaining relatively blind to the possibility of deception. Evolutionarily speaking, it is not surprising that agents have a built-in tendency to evaluate received testimony as honest.

4.4 The Limits of Vigilance

The dependence of vigilantism about testimony on inaccurate empirical assumptions means that the view fails. Because vigilantism underestimates the base rate of honesty and overestimates the effectiveness of monitoring, it overestimates the gains to be had by monitoring.

4.4.1 The Contribution of the Base Rate

Fricker's argument and the first version of Sperber's argument assume that the base rate of dishonesty is high (A1). We saw in Section 4.1 that this is not the case, and, as the modified Park-Levine model makes clear, the potential contribution of monitoring diminishes as the base rate of accurate representations increases. As we have seen, the reliability of testimonial belief-formation is given by the ratio of $P(H\&T)$ to $P(H\&T) + P(H\& \sim T)$. And $P(H\&T)$ and $P(H\& \sim T)$, in turn, depend on the base rate of honest testimony, $P(T)$.

As Park and Levine emphasize, it is misleading to say that deception detection accuracy is about 54 %, for this holds only when the base rate of honesty is 50 %; because subjects are strongly truth-biased, accuracy is strongly affected by the base rate. Similarly, as Fig. 1 shows, reliability of honesty judgements is strongly affected by the base rate, in such a way that there is little difference between the base rate and the reliability of honesty judgements—R is only slightly higher than $P(T)$. Thus there is little to be gained by monitoring and, at high base rates of honesty, accepting all or most received testimony will result in reliable testimonial belief-formation.

Thus, if the argument given above that the base rate of honesty is high is right, we can conclude that Fricker is wrong about the need for a reduction of testimonial justification. We can likewise conclude that the first version of Sperber's argument should be rejected: if deception is infrequent, agents need not effectively monitor for dishonesty in order for communication to be an evolutionarily stable strategy. Levine puts the point nicely: “evolving a finely tuned cognitive system adept at spotting leakage need not be an evolutionary mandate just because we get duped once in a while” (Levine 2010, p. 55).

4.4.2 The Contribution of Monitoring

As far as the epistemology is concerned, it is good news that the base rate of lying is low, for our poor ability to detect deception (contra A3) means that a reduction of testimonial justification is unavailable (agents are not able to filter out dishonest testimony, whether or not A2 is right)—Fricker’s argument fails. For the same reason, the first version of Sperber’s argument fails.

Our poor deception detection ability means that the success of the second version of Sperber’s argument, in contrast, will turn on the costs and benefits of a policy of epistemic vigilance relative to the alternatives: given the slight difference between R and $P(T)$, A3* seems to be right (presumably because A2* is right); the question, then, is whether, as A4 claims, this slight benefit of monitoring outweighs its costs.

Feasible policies for response to testimony can plausibly be ordered in terms of how costly they are as follows, from least costly to most costly.

Automatic rejection The cheapest policy is universal rejection, since it requires not only no resources for monitoring for dishonesty but also no resources for interpretation of testimony. Since this policy would deprive the agent of all testimonial information, it is not a realistic policy—the savings in terms of cognitive cost would be outweighed by the loss of access to information communicated by other agents. Thus we do not use this policy.

Automatic acceptance A somewhat more costly but still relatively cheap policy is automatic acceptance, or blind trust. While this policy, like the remaining policies, requires resources for interpretation of testimony, it requires no resources for monitoring. The policy gives the agent access to information communicated by other agents but leaves him entirely open to manipulation through deception. Since it cannot account for the slight difference between $P(T)$ and R (reliability will be the same as the base rate of accuracy—see Fig. 1), this policy, too, can be ruled out.

Default acceptance A more costly policy is default acceptance, with monitoring triggered only in special circumstances—when something about the current context (including prior knowledge or easily-noticeable aspects of the communicator’s behaviour) gives the agent reason to think that the communicator is likely to attempt to deceive him. This policy only occasionally requires resources for monitoring. The policy gives the agent access to communicated information and in principle provides him with some protection against manipulation through deception. How good this protection is in practice will depend, first, on the reliability with which context triggers monitoring on appropriate occasions (how likely it is that monitoring is triggered when the communicator is actually going to lie) and, second, on the reliability of contextually-triggered monitoring. It seems safe to assume that such contextually-triggered monitoring will be a type 2 process.

Default monitoring The most costly of the feasible policies is the default monitoring policy, since this always requires resources for monitoring.[10] The policy gives the agent access to communicated information and in principle provides him with protection against manipulation through deception. How good this protection is, depends, crucially, on the reliability of default monitoring. I assume that default monitoring is a type 1 process, though contextually-triggered type 2 monitoring will also be available to an agent employing the default monitoring policy.

Excluding the automatic rejection and automatic acceptance policies, what matters is the cost-benefit ratios of the remaining policies. The question is whether the difference between $P(T)$ and R is to be explained as the result of a default monitoring policy or, rather, a policy of default acceptance with contextually-triggered monitoring.

While Sperber's argument for default monitoring is tempting, there is reason to reject it: given that the cues to deception are weak and few, the task of detecting deception is difficult; thus the expected additional benefit of a default monitoring policy, relative to a default acceptance policy, is minimal, and is likely outweighed by the cognitive costs of monitoring. Given that most received messages are honest, it is adaptive to tend to simply assume that received messages are honest, rather than to waste resources in an attempt to determine whether they are honest. Moreover, default acceptance is superior to default monitoring in terms of speed (since it requires fewer resources) and power (since it results more often in formation of testimonial beliefs).

But can the contextually-triggered monitoring permitted by the default acceptance policy account for the difference between the base rate and our actual reliability? It appears likely that the policy can in fact account for this. With an eye to explaining the lack of variation in detection accuracy between individuals and across studies, Levine (2010) argues that there are "a few transparent liars", liars whose behaviour makes it so easy to determine that they are lying that the recipient can easily do this, while most communicators are not transparent liars. (Transparency can be affected by situational factors, not only communicator ability, so that the same communicator might be more or less transparent in different contexts.) This suggests that we can explain the fact that we do slightly better than chance at detecting deception as the

[10] Sperber et al. (2010) argue that some of the information required for detection of deception is necessarily acquired in the course of interpretation of communication. But this only goes so far in cutting down the cost of monitoring for dishonesty—meaningful monitoring will clearly require cognitive resources beyond those required for mere comprehension of an utterance. And even where the information is available, resources will still be required to do something with it.

result not of always-on monitoring but rather as the result of monitoring that is occasionally triggered by the unusual behaviour of the communicator. Of course, this means that recipients need to have some degree of sensitivity to the cues provided by transparent liars, but this is compatible with the default acceptance policy. The behaviour of transparent liars can make it apparent to receivers that they are lying, even if receivers do not employ a monitoring process dedicated to detecting cues: there need be no system 1 monitoring process employed by default; instead, unusual behaviour by the communicator induces the receiver to consciously monitor her; and this system 2 monitoring, once engaged, sometimes permits lie-detection. This proposal is broadly consistent with Thagard's default-and-trigger model of testimonial belief-formation (Thagard 2006), which distinguishes a default pathway of automatic acceptance of communicated information and a reflective pathway of reflective evaluation based on explanatory coherence; the default pathway is used unless incoherence of the content or lack of credibility of the source triggers use of the reflective pathway.

It thus seems likely that A4 is incorrect, and thus that the second version of Sperber's argument also fails. If so, vigilantism about testimony is untenable: monitoring cannot play the sort of role assigned to it by Fricker and Sperber.

5 Towards Vigilantism About Memory

It is intuitively plausible that testimony, as an external source, requires monitoring by the agent, since communicators will often have an incentive to deceive the agent, while memory, as an internal source, does not require such monitoring, since it is designed to serve the agent's interests. We have seen that the first half of this line is mistaken: a stance of default monitoring is neither normatively appropriate nor actually employed by agents with respect to testimony. In this section, focussing on episodic memory, I argue that the second half of the line is also mistaken: while memory is indeed designed to serve the agent's interests, this does not mean that it functions to provide the agent with mostly accurate information; in fact, it provides the agent with a great deal of inaccurate information, and, consequently, a stance of default monitoring is both normatively appropriate and actually employed by agents with respect to memory. In other words, while vigilantism is incorrect with respect to testimony, a form of vigilantism is correct with respect to memory.

5.1 Endorsement Problems

According to a simple, intuitively plausible picture of the operation of memory,[11] records are placed in memory when the subject endorses (believes)

[11] I draw here on Clark and Chalmers' paper on the extended mind hypothesis (Clark and Chalmers 1998), discussed more fully in Michaelian (2012d), but this sort of picture is implicit in many philosophical discussions of memory (e.g., Burge 1993).

them; records are discrete, stable items, remaining unchanged while in memory and unchanged during retrieval; and records are endorsed (believed) automatically upon retrieval. On this "preservationist" view (not to be confused with preservationism about memory justification (Lackey 2008)), the function of memory is simply to preserve the agent's beliefs.

If preservationism is right, then, assuming that memory performs its function well, vigilantism is clearly incorrect with respect to memory, both descriptively and normatively. Descriptively: assuming that the agent's other belief-forming processes are reliable, there is no advantage to be gained by monitoring one's own memory for signs of "deception". Normatively: making the same assumption, the agent will be entitled simply to trust information received from memory, for doing so is a reliable belief-forming (or belief-preserving) process. The problem is that preservationism is false. (I here rely on work done elsewhere (Michaelian 2011a, 2012d)—space does not permit reviewing my argument against preservationism in detail).

First, memory does not store only endorsed representations, and representations are not endorsed automatically at retrieval. Storage is determined by a form of relevance (where this covers a range of factors, including depth of processing Craik 2002), with the consequence that relevant but non-endorsed records are stored. As non-endorsed representations are stored, representations are not endorsed automatically at retrieval: endorsement at retrieval is determined by a range of metamemory processes (Michaelian 2012c; Hertwig et al. 2008; Mitchell and Johnson 2009). The key point is that retrieval is a two-level process—memory is a metacognitive belief-producing system. In terms of MEF: the memory store is the information source; metamemory processes monitor retrieval (process and retrieved content) and determine endorsement/rejection.

Second, a form of metacognition is required also due to the constructive character of memory. Memory records are not discrete, stable items but rather are transformed by a variety of broadly inferential processes during encoding, consolidation, retrieval, and reconsolidation following retrieval (Dudai 2004; Koriat et al. 2000; Loftus 2005; McClelland 2011; Schacter and Addis 2007). The upshot is that retrieval from memory can in fact mean the production of a new representation, a representation that was not previously stored in memory or even previously entertained by the agent. Thus the agent requires some means of determining whether to endorse this newly-produced representation.

Indeed, while it was for a long time standard to refer to a dedicated episodic memory system (Michaelian 2011b), the current tendency is to view episodic remembering rather as one function of a system capable of engaging in a broader range of constructive functions—either of a system devoted to "mental time travel" into both past and future (Tulving 1993) or, more radically, of a more general construction system (Hassabis and Maguire 2009). Even on the MTT approach, there is reason to take imagination of future events to be the primary function of the system: as Suddendorf and Corbalis argue, "our ability to revisit the past may be only a design feature of our ability to conceive of the future" (Suddendorf 2007, p. 303). And there is evidence (Spreng et al.

2009) that episodic remembering should be viewed as one activity of a system devoted not only to mental time travel but to a range of forms of imaginative construction of possible situations. Hassabis and Maguire (2009) argue that episodic memory is one of a number of different forms of "scene construction", including not only episodic future thinking but also navigation, theory of mind, mind wandering, and imagining fictitious experiences, which rely on the same brain network. The key point, for present purposes, is that on either the MTT hypothesis or the construction system hypothesis, the same system is responsible for the production of representations not only of the agent's past experiences but also of other possible experiences. Thus it is not only due to storage of non-endorsed information that remembering requires metacognition, but also because the agent requires some means of determining whether he is remembering or, rather, engaged in some other form of construction.

Thus, agents require some means of determining when to trust internally-generated representations, both because remembering must be distinguished from other constructive processes (including imagination) and because memories originating in experience must be distinguished from memories originating in other sources (imagination, etc.). As Urmson points out (Urmson 1967) (see also Bernecker 2008), it is important to distinguish these two questions: How do we manage to determine whether we are remembering rather than imagining? How do we manage to determine whether we are remembering successfully (accurately) rather than unsuccessfully? Distinguishing these questions, we can see that the remembering agent faces a double "endorsement problem": (1) the agent faces the task of distinguishing between remembering and other, related constructive processes—the *process problem*; (2) the agent faces the task of distinguishing between memories originating in experience and memories originating in other sources—the *source problem*.[12] Reliable formation of memory beliefs presupposes solving both the source problem and the process problem.

We can see the need for an explanation of how the agent manages to perform these tasks because, while we reliably manage to determine both the source of remembered information and whether we are remembering rather than engaging in some other form of construction, failure sometimes occurs for each task. My claim is that two similar but distinct forms of metacognition are responsible for solving these related problems: *process monitoring* allows the agent to determine whether she is remembering or engaging in some other form of construction; *source monitoring* allows the agent to distinguish between memories originating in experience and memories originating in other sources.

Source monitoring failures can, e.g., account for the misinformation effect (Michaelian 2012b; Loftus 2005; Lindsay 1994), in which post-event information is incorporated into the agent's memory representation of a witnessed

[12] In both the source problem and the process problem, there might be intermediate/indeterminate cases; I set these aside here.

event. Similarly, it is plausible that source monitoring failures account for some cases of imagination inflation, in which imagining an event increases confidence that it occurred (Garry et al. 1996).

The possibility of process monitoring failures was already noted by Hume:

> And as an idea of the memory, by losing its force and vivacity, may degenerate to such a degree, as to be taken for an idea of the imagination; so on the other hand an idea of the imagination may acquire such force and vivacity, so as to pass for an idea of the memory. (Hume 1739)

Note that, read in the most straightforward manner, Hume does not seem to be saying that an agent can take himself to be remembering an experience when in fact he is remembering something imagined (and vice versa), but rather that an agent can take himself to be remembering when in fact he is imagining (and vice versa). Failures of process monitoring resulting in taking imagining for remembering might be involved in such phenomena as delusions (Currie and Ravenscroft 2002), false recovered memories (Lindsay and Don Read 2005; Johnson et al. 2012), or discovery misattribution, in which the experience of solving a problem is confused with remembering (Dougal and Schooler 2007). Failures going in the other direction might be behind phenomena such as cryptomnesia (Brown and Murphy 1989; Marsh et al. 1997; Brédart et al. 2003). Regardless of the details, the occurrence of both types of error make it clear that there is a fallible mechanism at work in process monitoring just as much as in source monitoring, that the type of the process that is unfolding is not automatically and transparently given to the agent as part of its unfolding.

5.2 The Role of Metacognition: Process Monitoring

The philosophical literature on memory and imagination contains a number of suggestions that can be read as proposals about how agents solve the process problem;[13] these can be conveniently grouped into formal, content-based, and phenomenological solutions.

5.2.1 Formal Solutions

Formal solutions to the process problem claim that remembering is distinguished from imagining on the basis of structural features, either of the processes themselves or of the representations that they produce.

Flexibility In addition to his content-based solution (see Section 5.2.2), Hume (1739) suggests that memory and imagination can be distinguished on the basis of their relative flexiblity; the suggestion is, roughly, that whereas imagination flexibly recombines aspects of experience, memory is bound to preserve their

[13] I draw here on Bernecker's helpful discussion of memory markers in Bernecker (2008), noting where my approach overlaps with his.

original arrangement. However, as Bernecker points out (Bernecker 2008), the flexibility criterion is not usable, since the original experience cannot be compared to the present representation. Perhaps more seriously, the flexibility criterion fails to take the recontructive nature of remembering (Section 5.1 above) into account.

Intention On the approach advocated by Urmson (1967), the goal set by the agent determines the nature of the process: if the agent's goal in constructing a representation is to produce a representation of a past experience, then he is remembering; otherwise (if the construction is not so constrained), he is imagining. As Urmson puts it, "[w]e can infallibly determine whether we are recollecting or imagining simply by choosing" (Urmson 1967, p. 89–90). This proposal faces a number of problems. First, it makes error essentially impossible. On this proposal, unless the agent has a mistaken belief about what she wants to do, she cannot be in error about whether she is remembering or imagining. But presumably error is possible even where the agent does not have a mistaken belief about what she wants to do. Consider the case discussed by Martin and Deutscher of an agent who paints a scene that he takes himself to be imagining but who is mistaken about this—in fact, he is remembering a scene that he saw long ago, though he does not know that he is doing so (Martin and Deutscher 1966). Second, it ignores involuntary remembering. Remembering is not always (or even usually) voluntary (Hintzman 2011). On Urmson's proposal, the subject cannot know whether she is remembering or imagining unless she has first decided whether she is remembering or imagining. But it seems that the subject should be able to determine what she is doing even in cases of involuntary remembering. Finally, it overintellectualizes, by requiring the subject to have criteria for success in mind.

Voluntariness Furlong (1948, 1951) develops an approach according to which memory and imagination can be distinguished by their relative voluntariness: basically, while imagining is voluntary, memory is said to be involuntary. However, this proposal ignores both voluntary remembering and much mind wandering, which can be seen as a form of involuntary imagining.

5.2.2 Content-Based Solutions

Hume's other proposal for how the process problem is solved is that remembering is distinguished from imagining by its greater force and vivacity (Hume 1739). Though Bernecker refers to this as a "phenomenal" criterion, he reads the proposal as referring essentially to a supposed difference in the level of detail in the representations produced respectively by memory and imagination (Bernecker 2008). A similar proposal, on which memory can be distinguished from imagination by its greater level of contextual information, is entertained by Russell (1921). However, as Sutton points out (Sutton 1998), Hume himself undermines this proposal by pointing out that memory need not involve greater force and vivacity; nor is it clear that memory on average involves greater force and vivacity. Additionally, the content-based proposal does not appear to be able to distinguish between mental time travel into the past and mental time travel into the future (a point already discussed by Reid 1764), since the level of detail in representations of past and future events varies in a similar way. Finally, the content-based solution seems unlikely to be able to account for cases of what we might term "embedded construction"—level of detail seems unlikely to be able to distinguish among remembering remembering, imagining remembering, and so on (Bernecker 2008).

5.2.3 Phenomenological Solutions

A proposal modelled on the source monitoring framework (see Section 5.3) avoids these problems for formal and content-based approaches. Hume remarks that "[a]n idea assented to *feels* different from a fictitious idea, that the fancy alone presents to us" (Hume 1739). If we take Hume literally here, the suggestion would seem to be that, rather than being distinguished by their structure or content, remembering and imagining are distinguished by the different feelings that accompany the processes. Similar suggestions concerning phenomenological differences between remembering and imagining are made by a number of other theorists. James (1890) refers to feelings of warmth, intimacy, and the past direction of time. Russell (1921) refers to feelings of pastness and familiarity; similarly, a feeling of familiarity is referred to by Broad (1925), while Plantinga (1993) refers to a feeling of pastness.[14]

Developing the phenomenological solution, my basic proposal is that the type of the constructive cognitive process is determined using heuristics that go either from properties of the representation produced by the process or from phenomenal features of the process itself to judgements about the type of the process. It is implausible that process monitoring can rely on content alone, simply because imagination is often used to produce representations that are indistinguishable from those produced by memory, as far as their content is concerned, both intrinsically (level of detail, etc.) and in terms of their relation to other representations (coherence). Source monitoring, in contrast, can avoid this difficulty while relying entirely on features of content because it is concerned with distinguishing memory for imagination from memory for experience, and because the memory of an imagined representation will normally include information about the cognitive operations responsible for its production; but the representation produced by imagination itself does not include such information, so process monitoring cannot exploit this. It is thus more likely that process monitoring relies (primarily) on phenomenology.

I will not try to develop this suggestion in detail here, but it is plausible that there are phenomenal differences between the various forms of construction,

[14] Making a different sort of phenomenological proposal, Audi (1995) suggests that remembering is distinguished from imagining by a feeling of having believed; since memory both stores non-endorsed representations and is capable of producing new representations and beliefs, however, this proposal is a non-starter.

and that these can be used to determine (reliably, but not infallibly) the type of construction to which a given process belongs. One way this might work: In contrast to the mere episodic-like or what-where-when memory possessed by some animals (Suddendorf 2007; Roberts and Feeney 2009), true episodic memory involves a distinctive phenomenology, a sense of subjective time (Tulving 1983; Nyberg et al. 2010). This phenomenology is shared by mental time travel into to the future (Boyer 2008). So episodic remembering can be distinguished from episodic future thinking in terms of the directionality of the sense of subjective time that they involve—I have the sense that this is something that has happened to me, or I have the sense that this is something that will happen to me. And imagination can be distinguished from MTT in general because it does not normally involve the same sort of sense of subjective time. These discriminations are presumably accomplished by type 1 processes, so that the agent is not aware of making them. A complete story, of course, will have to explain also how the agent can distinguish between memory or imagination and other forms of construction; my aim here is, modestly, to make plausible the suggestion that, drawing on empirical work on phenomenological differences between remembering and imagining (D'Argembeau and Van der Linden 2004, 2006; Gamboz et al. 2010; de Vito et al. 2012), a process monitoring framework analogous to the source monitoring framework can be developed.

5.3 The Role of Metacognition: Source Monitoring

The suggestion that source monitoring solves the second half of the endorsement problem is similarly speculative, though better supported. Because we cannot establish the base rate of accurate representations produced by memory ($P(T)$) with much precision, and because we similarly cannot establish the reliability of source monitoring ($P(H|T)$ and $P(\sim H| \sim T)$) with much precision, we cannot determine the reliability of remembering with any precision. But since the assumption that memory is reliable is safe (Michaelian 2012b), and since the source monitoring framework, which is itself grounded in a developed body of theory and research on metamemory processes (see Mitchell and Johnson 2009, for a recent review), was designed in part precisely to provide an explanation of how memory can be reliable despite the poor reliability of retrieval, the framework does suggest that source monitoring solves the problem.

Source monitoring theorists argue that we are able to discriminate the origins of mental experiences by means of attributional judgements processes, evaluative or monitoring processes which take us from properties of retrieved information (and, in certain cases, features of its relation to other memories) to a judgement that the information stems from a certain source (and thus is or is not likely to be veridical) (Mitchell and Johnson 2000, p. 180). Though memory does not normally store information about source, memories typically bear characteristic marks of the sources in which they originate; for example, "memories of imagined events typically have less vivid perceptual, temporal, and spatial information than perceived events and often include information about intentional cognitive operations …”, while “[m]emories of dreams are often perceptually vivid, typically do not include information about the cognitive operations that created them, and are often inconsistent with knowledge or other memories” (Mitchell and Johnson 2000, p. 180). The presence of these marks means that though information about the source of a record is not typically stored along with the record itself, it is nevertheless possible to determine the source of a record with some reliability, using “heuristic source monitoring processes to attribute a source to information based on an evaluation of various features of the information” (Johnson and Raye 2000, p. 39).

Source monitoring is normally a type 1 process, unconscious and automatic (Mitchell and Johnson 2000; Johnson and Raye 2000), though it can also be performed by type 2 processes. Thus, in terms of MEF, the source monitoring framework suggests that the endorsement policy that agents employ with respect to their own memories is analogous to the default monitoring policy that vigilantists incorrectly take them to employ with respect to testimony.

5.4 The Contributions of the Base Rate and Monitoring in Memory

Given the points about endorsement and construction, we are entitled to assumptions analogous to those required by Fricker and Sperber (on the first reading):

A1-M Inaccurate memories are frequent, as are representations produced by imagination but which could be mistaken for memories, so that accepting apparent memories without monitoring would lead to unreliable belief-formation.

If the source monitoring framework and the process monitoring framework are on the right track, then we are also entitled to assume:

A2-M Reliable cues to accuracy (to whether the representation is a memory and, if so, which source it originates in) exist.

A3-M Subjects can and do exploit these cues to monitor effectively for inaccuracy.

These assumptions ground a form of vigilantism about memory. Descriptively: in order for memory to be sufficiently reliable to be beneficial, agents must be epistemically vigilant with respect to remembered information. Normatively: We are not entitled simply to trust our memories but must monitor memory for accuracy.

6 Conclusions

One might object that it is a mistake to oppose trust to vigilance, as I have done here, that, rather than being opposed to vigilance, trust is in fact *based on* vigilance. Drawing on work by Origgi (2004, 2010, 2012), Sperber et al. suggest precisely this sort of view, which they illustrate using the following analogy:

> 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. (Sperber et al. 2010, 364)

While the thought that trust is based on vigilance is seductive, I believe that this analogy does not do the work that Sperber et al. want it to do.

In a crowd of pedestrians, each person is normally trying to avoid colliding with others—their interests, as far as collisions are concerned, coincide. But the appeal of vigilantism about testimony relies precisely on the fact that, in communication, the interests of communicator and receiver often do not coincide: whereas it is unusual for someone to attempt to collide with you, it is supposed to be an ordinary or even a frequent occurrence for someone to attempt to deceive you. Vigilance with respect to pedestrian collisions when walking thus is not analogous to vigilance with respect to deception in communication in the manner that Sperber et al. suggest. In the pedestrian case, it makes sense to say that trust is based on vigilance: we trust each other because each of us knows that the others are vigilant; but what it is to be vigilant in this context is precisely to try not to collide. There is nothing analogous in the testimony case, as vigilantism conceives of it: there, each of us must be vigilant precisely because others are trying to "collide" with us; here, trust is not based on vigilance.

The analogy can be reworked: The vigilance we exercise in communication is, if the approach developed here is right, indeed analogous to the vigilance we exercise when walking in a crowd of pedestrians, but not in the way that Sperber et al. suggest. When we walk down the street, we mostly trust others not to collide with us, just as we mostly try not to collide with others; we do not need to devote significant resources to monitoring for others who are attempting to collide with us, but rather concern ourselves with controlling our own trajectory so that we do not collide with them. Occasionally, however, unusual behaviour causes us to consciously take deliberate steps to avoid a collision—when someone appears to be drunk or aggressive, for example, we might begin to monitor his trajectory in order to determine whether he is likely to collide with us. When this happens, we can usually avoid the potential collision, by withdrawing our default trust and steering clear of him if necessary. Things are, I claim, similar in communication. When we exchange information with others, we mostly trust them to tell us the truth, just as we mostly try to tell the truth to them; we do not need to devote significant resources to monitoring for attempts to deceive us, but rather concern ourselves with giving accurate information when we testify. Occasionally, however, unusual behaviour causes us to consciously take deliberate steps to avoid being deceived—when someone is suddenly extremely nervous, for example, we might begin to monitor him to attempt to determine whether he is lying to us. My suggestion is that in such situations, we usually manage to avoid being deceived: we withdraw our default trust in the communicator and attempt to determine whether he is indeed lying, and evaluations made under these circumstances are reliable.

Things are, I have argued, different as far as memory is concerned—we should invert the vigilantist line. This inversion is correct both descriptively and normatively: descriptively, because monitoring of testimony is limited and ineffective, while the construction system is monitored regularly and effectively; normatively, because, while social metacognition is relatively ineffective, individual metacognition is relatively effective. While counterintuitive, this inversion in the end should perhaps not be surprising, as agents will normally have much better information about their own cognitive operations than they do about those of others.

Acknowledgements Thanks for comments and discussion to Joëlle Proust, Sam Wilkinson, two anonymous referees, and audiences at the Third Copenhagen Conference in Epistemology (Social Epistemology Research Group, Københavns Universitet), a meeting of the Filosofiska Föreningen at Lunds Universitet (organized by Frank Zenker), and the "Brains, Minds, and Language" workshop at Bogaziçi Üniversitesi (organized by Lucas Thorpe and Ali Salah).

References

Alicke, Mark D., and Constantine Sedikides. 2009. Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology* 20:1–48.

Anderson, D. Eric, Bella M. DePaulo, and Matthew E. Ansfield. 2002. The development of deception detection skill: A longitudinal study of same-sex friends. *Personality and Social Psychology Bulletin* 28(4):536–545.

Anderson, Michael L., Tim Oates, Waiyian Chong, and Don Perlis. 2006. The metacognitive loop I: Enhancing reinforcement learning with metacognitive monitoring and control for improved perturbation tolerance. *Journal of Experimental & Theoretical Artificial Intelligence* 18(3):387–411.

Audi, R. 1995. Memorial justification. *Philosophical Topics* 23:31–45

Bernecker, S. 2008. *The metaphysics of memory*. Springer.

Bernecker, S. 2010. *Memory: A philosophical study*. Oxford: Oxford University Press.

Bond, Charles F., and Bella M. Depaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review* 10(3):214–234

Bond, Charles F., and Bella M. DePaulo. 2008. Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin* 134(4):477–492.

Boyer, P. 2008. Evolutionary economics of mental time travel? *Trends in Cognitive Sciences* 12(6):219–224.

Brédart, Serge, James Lampinen, and Anne-Catherine Defeldre. 2003. Phenomenal characteristics of cryptomnesia. *Memory* 11(1):1–11

Broad, C.D. 1925. *The mind and its place in nature*. New York: The Humanities Press.

Brown, Alan S., and Dana R. Murphy. 1989. Cryptomnesia: Delineating inadvertent plagiarism. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15(3):432–442.

Burge, Tyler. 1993. Content preservation. *The Philosophical Review* 102(4):457–488.

Christ, Shawn E., David C. Van Essen, Jason M. Watson, Lindsay E. Brubaker, and Kathleen B. McDermott. 2009. The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex (New York, N.Y.: 1991)* 19(7):1557–1566.

Clark, Andy, and David Chalmers. 1998. The extended mind. *Analysis* 58(1):7–19.

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

Cox, M. 2005. Metacognition in computation: A selected research review. *Artificial Intelligence* 169(2):104–141.

Craik, Fergus I. 2002. Levels of processing: Past, present ... and future? *Memory* 10(5/6):305–318.

Cummins, Robert, Pierre Poirier, and Martin Roth. 2004. Epistemological strata and the rules of right reason. *Synthese* 141(3):287–331.

Currie, G., and I. Ravenscroft. 2002. *Recreative minds*. Oxford: Oxford University Press.

D'Argembeau, Arnaud, and Martial Van der Linden. 2004. Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: Influence of valence and temporal distance. *Consciousness and Cognition* 13(4):844–858.

D'Argembeau, Arnaud, and Martial Van der Linden. 2006. Individual differences in the phenomenology of mental time travel: The effect of vivid visual imagery and emotion regulation strategies. *Consciousness and Cognition* 15(2):342–350.

de Sousa, R. 2008. Epistemic feelings. In *Epistemology and emotions*, eds. U. Doğuoğlu, and D. Kuenzle. Ashgate.

de Vito, Stefania, Nadia Gamboz, and Maria A. Brandimonte. 2012. What differentiates episodic future thinking from complex scene imagery? *Consciousness and Cognition* 21(2):812–813.

DePaulo, B.M., D.A. Kashy, S.E. Kirkendol, M.M. Wyer, and J.A. Epstein. 1996. Lying in everyday life. *Journal of Personality and Social Psychology* 70(5):979–995.

DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin* 129(1):74–118.

Dokic, J. 2013. Foundations of Metacognition. In *Metacognition, mental agency and self-awareness*, eds. M. Beran, J. Brandl, J. Perner, and J. Proust. Oxford: Oxford University Press (forthcoming).

Dougal, Sonya, and Jonathan W. Schooler. 2007. Discovery misattribution: When solving is confused with remembering. *Journal of Experimental Psychology: General*, 136(4):577–592.

Douven, Igor, and Stefaan E. Cuypers. 2009. Fricker on testimonial justification. *Studies in History and Philosophy of Science Part A* 40(1):36–44.

Dudai, Yadin. 2004. The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology* 55(1):51–86.

Dunlosky, J., and R.A. Bjork, eds. 2008. *Handbook of metamemory and memory*. New York: Psychology Press.

Evans, J.St.B.T. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59(1):255–278.

Forrest, James A., Robert S. Feldman, and James M. Tyler. 2004. When accurate beliefs lead to better lie detection1. *Journal of Applied Social Psychology* 34(4):764–780.

Frankish, Keith. 2010. Dual-process and dual-system theories of reasoning. *Philosophy Compass* 5(10):914–926.

Fricker, E. 1994. Against gullibility. In *Knowing from words*, eds. B.K. Matilal, and A. Chakrabarti, 125–161. Dordrecht: Kluwer.

Fricker, E. 2004. Testimony: Knowing through being told. In *Handbook of epistemology*, eds. I. Niiniluoto, M. Sintonen, and J. Woleński, 109–130. Dordrecht: Kluwer.

Fricker, E. 2006a. Testimony and epistemic autonomy. In *The epistemology of testimony*, eds. J. Lackey, and E. Sosa, 225–253. Oxford: Clarendon.

Fricker, Elizabeth. 1995. Telling and trusting: Reductionism and anti-reductionism in the epistemology of testimony. *Mind* 104(414):393–411.

Fricker, Elizabeth. 2002. Trusting others in the sciences: a priori or empirical warrant? *Studies in History and Philosophy of Science Part A* 33(2):373–383.

Fricker, Elizabeth. 2006b. Second-hand knowledge*. *Philosophy and Phenomenological Research* 73(3):592–618.

Fricker, Elizabeth. 2006c. Varieties of anti-reductionism about testimony-a reply to Goldberg and Henderson. *Philosophy and Phenomenological Research* 72(3):618–628.

Furlong, E.J. 1948. Memory. *Mind* 57(225):16–44.

Furlong, E.J. 1951. *A study in memory: A philosophical essay*.

Gamboz, Nadia, Maria A. Brandimonte, and Stefania De Vito. 2010. The role of past in the simulation of autobiographical future episodes. *Experimental Psychology* 57(6):419–428.

Garry, Maryanne, Charles Manning, Elizabeth Loftus, and Steven Sherman. 1996. Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin & Review* 3(2):208–214.

Gelfert, Axel. 2009. Indefensible middle ground for local reductionism about testimony. *Ratio* 22(2):170–190.

Gilbert, Daniel T. 1991. How mental systems believe. *American Psychologist* 46(2):107–119.

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

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

Goldman, Alvin. 1979. What is justified belief? In *Justification and knowledge: New studies in epistemology*, ed. George S. Pappas, 1–23. Dordrecht: Reidel (Reprinted in Goldman, Dordrecht. 1992. *Liaisons*. Cambridge: MIT Press).

Goldman, Dordrecht. 1992. *Liaisons*. Cambridge: MIT Press.

Grafen, A. 1990. Biological signals as handicaps. *Journal of Theoretical Biology* 144(4):517–546.

Grice, P. 1989. *Studies in the way of words*. Harvard University Press.

Hassabis, Demis, and Eleanor A. Maguire. 2009. The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1521):1263–1271.

Hasson, Uri, Joseph P. Simmons, and Alexander Todorov. 2005. Believe it or not: On the possibility of suspending belief. *Psychological Science: A Journal of the American Psychological Society/APS* 16(7):566–571.

Hertwig, Ralph, Stefan M. Herzog, Lael J. Schooler, and Torsten Reimer. 2008. Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 34(5):1191–1206.

Hintzman, Douglas L. 2011. Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth". *Perspectives on Psychological Science* 6(3):253–271.

Hume, David. 1739. *A treatise of human nature*.

Inglehart, R., M. Basanez, and A. Moreno, eds. 1998. *Human values and beliefs: A cross-cultural sourcebook*. University of Michigan Press.

Inglehart, R., M. Basanez, J. Diez-Medrano, L. Halman, and R. Luijkx, eds. 2004. *Human beliefs and values: A cross-cultural sourcebook based on the 1999–2002 values surveys*. Siglo XXI Editores.

James, William. 1890. *The Principles of Psychology*.

Johnson, M.K., and C.L. Raye. 2000. Cognitive and brain mechanisms of false memories and beliefs. In *Memory, brain, and belief*, eds. D.L. Schacter, and E. Scarry, 35–86. Harvard University Press, Cambridge.

Johnson, Marcia K., Carol L. Raye, Karen J. Mitchell, and Elizabeth Ankudowich. 2012. The cognitive neuroscience of true and false memories true and false recovered memories. In *Nebraska symposium on motivation*, chap 2, vol 58, pp 15–52. New York, NY: Springer New York.

Jost, John T., Arie W. Kruglanski, and Thomas O. Nelson. 1998. Social metacognition: An expansionist review. *Personality and Social Psychology Review* 2(2):137–154.

Koriat, A. 2006. Metacognition and consciousness. In *Cambridge handbook of consciousness*, eds. P.D. Zelazo, M. Moscovitch, and E. Thompson. Cambridge, UK: Cambridge University Press.

Koriat, Asher. 1998. Metamemory: The feeling of knowing and its vagaries. In *In psychological science, vol. 2: Biological and cognitive aspects*, eds. Michel Sabourin, Fergus Craik, and Michèle Robert. Psychology Press.

Koriat, A., M. Goldsmith, and A. Pansky. 2000. Toward a psychology of memory accuracy. *Annual Review of Psychology* 51:481–537.

Kraut, Robert. 1980. Humans as lie detectors. *Journal of Communication* 30(4):209–218.

Kumkale, G.T., and D. Albarracín. 2004. The sleeper effect in persuasion: a meta-analytic review. *Psychological Bulletin* 130(1):143–172.

Lackey, J. 2008. *Learning from words: Testimony as a source of knowledge*. Oxford: Oxford University Press.

Lepock, C. 2007. *Metacognition and intellectual virtue*. PhD thesis, University of Alberta, Edmonton.

Levin, D. 2002. Change blindness blindness: As visual metacognition. *Journal of Consciousness Studies* 9(5–6):111–130.

Levine, T.R. 2010. A few transparent liars: explaining 54 % accuracy in deception detection experiments. In *Communication yearbook 34*, ed. C.T. Salmon, 41–61. Routledge.

Levine, T.R., and R.K. Kim. 2010. Some considerations for a new theory of deceptive communication. In *The interplay of truth and deception*, eds. M. Knapp, and M. McGlone, 16–34. Routledge.

Levine, Timothy R., Hee S. Park, and Steven A. McCornack. 1999. Accuracy in detecting truths and lies: Documenting the veracity effect. *Communication Monographs* 66(2):125–144.

Levine, Timothy R., Rachel K. Kim, Hee S. Park, and Mikayla Hughes. 2006. Deception detection accuracy is a predictable linear function of message veracity base-rate: A formal test of Park and Levine's probability model. *Communication Monographs* 73(3):243–260.

Levine, Timothy R., Rachel K. Kim, and Lauren M. Hamel. 2010. People lie for a reason: Three experiments documenting the principle of veracity. *Communication Research Reports* 27(4):271–285.

Lindsay, D.S. 1994. Memory source monitoring and eyewitness testimony. In *Adult eyewitness testimony: Current trends and developments*, eds. D.F. Ross, J.D. Read, and M.P. Toglia. New York: Cambridge University Press.

Lindsay, D. Stephen, and J. Don Read. 2005. The recovered memories controversy: Where do we go from here? In *Recovered memories: Seeking the middle ground*, 69–93. Wiley.

Loftus, Elizabeth F. 2005. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory* 12(4):361–366.

Loussouarn, A. 2010. *De la métaperception à l'agir perceptif*. PhD thesis, Institut Jean-Nicod/Ecole des Hautes Etudes en Sciences Sociales, Paris.

Marsh, Richard L., Joshua D. Landau, and Jason L. Hicks. 1997. Contributions of inadequate source monitoring to unconscious plagiarism during idea generation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23(4):886–897.

Martin, C.B., and Max Deutscher. 1966. Remembering. *The Philosophical Review* 75(2):161–196.

Mascaro, Olivier, and Dan Sperber. 2009. The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition* 112(3):367–380.

Matthen, Mohan. 2010. Is memory preservation? *Philosophical Studies* 148(1):3–14.

McClelland, J.L. 2011. Memory as a constructive process: The parallel-distributed processing approach. In *The memory process: Neuroscientific and humanistic perspectives*, eds. S. Nalbantian, P. Matthews, and J.L. McClelland, 129–151. Cambridge, MA: MIT Press.

McCornack, Steven A., and Timothy R. Levine. 1990. When lovers become leery: The relationship between suspicion and accuracy in detecting deception. *Communication Monographs* 57(3):219–230.

McCornack, S.A., and M.R. Parks. 1986. Deception detection and relational development: The other side of trust. In *Communication yearbook 9*, ed. M.L. McLaughlin, 377–389. Sage.

McKay, Ryan T., and Daniel C. Dennett. 2009. The evolution of misbelief. *Behavioral and Brain Sciences* 32(6):493–510.

Metcalfe, J. 2008. Evolution of metacognition. In *Handbook of metamemory and memory*, eds. J. Dunlosky, and R.A. Bjork, 29–46. New York: Psychology Press.

Michaelian, Kourken. 2008. Testimony as a natural kind. *Episteme: A Journal of Social Epistemology* 5(2):180–202.

Michaelian, Kourken. 2010. In defence of gullibility: The epistemology of testimony and the psychology of deception detection. *Synthese* 176(3):399–427.

Michaelian, Kourken. 2011a. Generative memory. *Philosophical Psychology* 24(3):323–342.

Michaelian, Kourken. 2011b. Is memory a natural kind? *Memory Studies* 4(2):170–189.

Michaelian, Kourken. 2011c. The epistemology of forgetting. *Erkenntnis* 74(3):399–424.

Michaelian, K. 2012a. The evolution of testimony: Receiver vigilance, speaker honesty, and the reliability of communication. (forthcoming).

Michaelian, K. 2012b. The information effect: Constructive memory, testimony, and epistemic luck. *Synthese* (forthcoming).

Michaelian, K. 2012c. Metacognition and endorsement. *Mind & Language* 27(3):284–307.

Michaelian, Kourken. 2012d. Is external memory memory? Biological memory and extended mind. *Consciousness and Cognition* (forthcoming).

Millar, Murray, and Karen Millar. 1995. Detection of deception in familiar and unfamiliar persons: The effects of information restriction. *Journal of Nonverbal Behavior* 19(2):69–84.

Mitchell, K.J., and M.K. Johnson. 2000. Source monitoring: Attributing mental experiences. In *Oxford handbook of memory*, eds. E. Tulving, and F.I.M. Craik, 175–95. Oxford: Oxford University Press.

Mitchell, Karen J., and Marcia K. Johnson. 2009. Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin* 135(4):638–677.

Nelson, T.O., and L. Narens. 1994. Why investigate metacognition? In *Metacognition*, eds. J. Metcalfe, and A.P. Shimamura, 1–26. Cambridge: MIT Press.

Nyberg, Lars, Alice S. N. Kim, Reza Habib, Brian Levine, and Endel Tulving. 2010. Consciousness of subjective time in the brain. *Proceedings of the National Academy of Sciences* 107(51):22356–22359.

Oppenheimer, Daniel M. 2008. The secret life of fluency. *Trends in Cognitive Sciences* 12(6): 237–241.

Origgi, Gloria. 2004. Is trust an epistemological notion? *Episteme* 1(01):61–72.

Origgi, Gloria. 2010. Epistemic vigilance and epistemic responsibility in the liquid world of scientific publications. *Social Epistemology* 24(3):149–159.

Origgi, Gloria. 2012. Epistemic injustice and epistemic trust. *Social Epistemology* 26(2):221–235.

Park, Hee S., and Timothy Levine. 2001. A probability model of accuracy in deception detection experiments. *Communication Monographs* 68(2):201–210.

Plantinga, A. 1993. *Warrant and proper function*. Oxford.

Proust, J. 2006. Rationality and metacognition in non-human animals. In *Rational animals*, eds. S. Hurley, and M. Nudds, 247–274. Oxford: Oxford University Press.

Proust, J. 2008. Epistemic agency and metacognition: An externalist view. *Proceedings of the Aristotelian Society* 108:241–268.

Reid, T. 1764/1970. *An inquiry into the human mind*. Chicago: University of Chicago Press.

Richter, Tobias, Sascha Schroeder, and Britta Wöhrmann. 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(3):538–558.

Roberts, William A., and Miranda C. Feeney. 2009. The comparative study of mental time travel. *Trends in Cognitive Sciences* 13(6):271–277.

Russell, B. 1921. *The analysis of mind*. London: George Allen & Unwin.

Schacter, Daniel L., and Donna R. Addis. 2007. The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362(1481):773–786.

Scott-Phillips, Thomas C. 2008. On the correct application of animal signalling theory to human communication. In *Proceedings of the 7th international conference on the evolution of language*, eds. A.D.M. Smith, K. Smith, and R. Ferrer i Cancho, 275–282. Singapore: World Scientific.

Searcy, W.R., and S. Nowicki. 2005. *The evolution of animal communication: Reliability and deception in signalling systems*. Princeton, NJ: Princeton University Press.

Senor, Thomas D. 2007. Preserving preservationism: A reply to lackey. *Philosophy and Phenomenological Research* 74(1):199–208.

Serota, Kim B., Timothy R. Levine, and Franklin J. Boster. 2010. The prevalence of lying in America: Three studies of self-reported lies. *Human Communication Research* 36(1):2–25.

Smith, J.D., W.E. Shields, and D.A. Washburn. 2003. The comparative psychology of uncertainty monitoring and metacognition. *Behavioural Brain Research* 26(3):317–373.

Sosa, E. 2007. *A virtue epistemology*. Oxford: Clarendon Press.

Spence, Sean A., and Catherine J. Kaylor-Hughes. 2008. Looking for truth and finding lies: The prospects for a nascent neuroimaging of deception. *Neurocase* 14(1):68–81.

Sperber, D. 2001. An evolutionary perspective on testimony and argumentation. *Philosophical Topics* 29:401–413.

Sperber, D., F. Clément, C. Heintz, O. Mascaro, H. Mercier, G. Origgi, and D. Wilson. 2010. Epistemic vigilance. *Mind & Language* 25(4):359–393.

Spreng, R. Nathan, Raymond A. Mar, and Alice S. Kim. 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience* 21(3):489–510.

Suddendorf, Thomas, and Michael C. Corballis. 2007. The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences* 30(3):299–313.

Sutton, J. 1998. *Philosophy and memory traces*. Cambridge: Cambridge.

The Global Deception Research Team. 2006. A world of lies. *Journal of Cross-Cultural Psychology* 37(1):60–74.

Thagard, Paul. 2006. Testimony, credibility, and explanatory coherence. *Erkenntnis* 63(3):295–316.

Thompson, Valerie. 2010. Towards a metacognitive dual process theory of conditional reasoning. In *Cognition and conditions*, ed. Mike Oaksford and Nick Chater. Oxford: Oxford University Press.

Tulving, E. 1983. *Elements of episodic memory*. Oxford: Oxford University Press.

Tulving, Endel. 1993. What is episodic memory? *Current Directions in Psychological Science* 2(3):67–70.

Urmson, J.O. 1967. Memory and imagination. *Mind* 76(301):83–91.

Verschuere, Bruno, Adriaan Spruyt, Ewout H. Meijer, and Henry Otgaar. 2011. The ease of lying. *Consciousness and Cognition* 20(3):908–911.

von Hippel, William, and Robert Trivers. 2011. The evolution and psychology of self-deception. *Behavioral and Brain Sciences* 34(1):1–16.

Vrij, A. 2000. *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. New York: Wiley.

Vrij, A. 2008. *Detecting lies and deceit: Pitfalls and opportunities*, 2nd edn. New York: Wiley.

Vrij, Aldert, Ronald Fisher, Samantha Mann, and Sharon Leal. 2006. Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences* 10(4):141–142.

Vrij, Aldert, Samantha Mann, Ronald Fisher, Sharon Leal, Rebecca Milne, and Ray Bull. 2008. Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior* 32(3):253–265.

Vrij, Aldert, Pär A. Granhag, Samantha Mann, and Sharon Leal. 2011. Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science* 20(1): 28–32.

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