Episteme, 10, 1 (2013) 37–59 © Cambridge University Press
doi:10.1017/epi.2013.2

THE EVOLUTION OF TESTIMONY: RECEIVER VIGILANCE, SPEAKER HONESTY AND THE RELIABILITY OF COMMUNICATION

KOURKEN MICHAELIAN
kmichaelian@bilkent.edu.tr

ABSTRACT

Drawing on both empirical evidence and evolutionary considerations, Sperber et al. argue that humans have a suite of evolved mechanisms for 'epistemic vigilance'. On their view, vigilance plays a crucial role in ensuring the reliability and hence the evolutionary stability of communication. This article responds to their argument for vigilance, drawing on additional empirical evidence (from deception detection research) and evolutionary considerations (from animal signalling research) to defend a more optimistic, quasi-Reidian view of communication. On this alternative view, the lion's share of the responsibility for explaining the reliability of testimony falls not to the vigilance of receivers but rather to the honesty of communicators, implying that vigilance does not play a major role in explaining the evolutionary stability of communication.

Drawing on both empirical evidence and evolutionary considerations, Sperber et al. argue that humans have a suite of evolved mechanisms for 'epistemic vigilance'. On their view, vigilance plays a crucial role in ensuring the reliability and hence the evolutionary stability of communication. This article responds to their argument for vigilance, drawing on additional empirical evidence (from deception detection research) and evolutionary considerations (from animal signalling research) to defend a more optimistic, quasi-Reidian view of communication. On this alternative view, the lion's share of the responsibility for explaining the reliability of testimony falls not to the vigilance of receivers but rather to the honesty of communicators, implying that vigilance does not play a major role in explaining the evolutionary stability of communication.

Section 1 reviews Sperber et al.'s (2010) case for vigilance, distinguishing two different readings of the argument. On the first, vigilance is indispensable to the reliability of communication; on the second, while vigilance is not indispensable, it is adaptive. Section 2 then distinguishes among several forms of vigilance. The first reading of Sperber et al.'s argument supports a strongly effective form of type 1 vigilance, while the second supports a moderately effective form of type 1 vigilance. In the remainder of the paper, I argue that we should accept only a weakly effective form of type 2 vigilance. Section 3 argues that Sperber et al.'s evidence does not support effective vigilance, and that there is considerable evidence that vigilance is neither strongly nor moderately effective; it appears, however, that vigilance is weakly effective. Section 4 concludes by arguing on evolutionary grounds that weakly effective vigilance should be viewed in terms of type 2 rather than type 1

processes. The overall picture that emerges is one on which the reliability of communication is ensured largely by the prevalence of honest communication, with type 2 vigilance making a minor contribution to reliability by filtering out some dishonest communicated information.

1. THE EVOLUTION OF VIGILANCE

Two readings of Sperber et al.'s (2010) evolutionary argument can be distinguished. On either reading, the argument supports the conclusion that '[a] disposition to be vigilant is likely to have evolved biologically alongside the ability to communicate in the way that humans do' (Sperber et al. 2010: 360), but they require somewhat different assumptions.

1.1 Vigilance as indispensable

As is standard, Sperber et al. conceive of a speaker's trustworthiness as having two components: competence and honesty. As they point out, if we assume that agents are largely competent, the central question about vigilance concerns vigilance with respect to honesty:

> [T]he major problem posed by communicated information has to do not with the competence of others, but with their interests and their honesty. While the interests of others often overlap with our own, they rarely coincide with ours exactly. In a variety of situations, their interests are best served by misleading or deceiving us. It is because of the risk of deception that epistemic vigilance may be not merely advantageous but indispensable if communication itself is to remain advantageous. (2010: 359–60)

I focus here on the role of vigilance with respect to honesty in explaining the reliability of communication, arguing that such vigilance makes only a minor contribution to reliability.

This leaves open the role played by vigilance with respect to competence in explaining the reliability of communication. If such vigilance turns out to be more effective than is vigilance with respect to honesty, it may be that Sperber et al. are right in assigning a central role to vigilance in explaining the reliability of communication. I suspect that this sort of view of the contribution of vigilance with respect to competence is not workable. First, as Sperber et al. themselves point out, '[i]n general, others are mistaken no more often than we are ... and they know things that we don't know. So it should be advantageous to rely even blindly on the competence of others' (2010: 359). This makes it plausible that the contribution of vigilance with respect to competence is relatively minor, though whether such vigilance is nevertheless adaptive will turn on its costs and benefits. Second, the view would require the assumption that the base rate of incompetent testimony is high (this is analogous to the assumption about the base rate of dishonest testimony discussed in section 4). There is reason to doubt this assumption: the view that agents are generally competent is plausible, since it is equivalent to the assumption that non-testimonial belief-formation is reliable, which is supported by the point that false belief is in general maladaptive.[1] But I do not attempt to develop these considerations into an argument here.

1 This sort of evolutionary argument for reliability is fairly standard, though it can be challenged. See e.g. McKay and Dennett 2009 for a recent discussion, and Stich 1990 for a challenge to such evolutionary arguments. The question of the effectiveness of vigilance relative to dishonesty is of interest even if one rejects the assumption of general competence.

Sperber et al.'s evolutionary reasoning for the claim that incentives for dishonesty necessitate vigilance expands an argument given earlier by Sperber. Against an optimistic view (associated with Reidian credulism (Coady 1992) in the epistemology of testimony and analogous to the classical ethological view of animal signalling (Smith 1977)) on which the function of communication is essentially to transmit accurate information, Sperber argues that '[c]ommunication produces a certain amount of misinformation in the performance of its function, more specifically in the performance of those aspects of its function that are beneficial to the communicator' (2001). While, from the point of view of the receiver, testimony functions to provide accurate information, from the point of the communicator, it functions to influence the behaviour of the receiver in ways that are beneficial to the communicator. And while desired behaviour can sometimes be produced by providing accurate information, it is often the case that providing inaccurate information is the best means of producing the desired behaviour.

Thus, in a move analogous to that influentially made by Dawkins and Krebs in the field of animal signalling (Dawkins and Krebs 1978; Krebs and Dawkins 1984), Sperber foregrounds the fact that communicators regularly have an incentive to deceive receivers. The resulting tendency of communicators to attempt to deceive gives rise to epistemic vigilance, for

> [i]f communication were on the whole beneficial to producers of messages ... at the expense of receivers, or beneficial to receivers at the expense of producers, one of the two behaviors would be likely to have been selected out, and the other behavior would have collapsed by the same token. In other words, for communication to evolve, it must be a positive-sum game where, in the long run at least, both communicators and receivers stand to gain. (2001)

In the case of animal signalling, Zahavi's handicap principle (Zahavi and Zahavi 1997) is usually invoked to account for the reliability of communication. According to the principle, signals are reliable because they are costly and therefore hard to fake. (Because the peacock's tail entails significant costs, only high-quality peacocks can produce impressive tails, so an impressive tail is a reliable signal of the peacock's quality as a mate.) But the production of a testimonial utterance is normally cheap (lying is easy), so handicaps cannot plausibly be invoked as the mechanism that accounts for the reliability of human communication. Sperber concludes that, 'if communication has stabilized among humans, it must be that there are ways to calibrate one's confidence in communicated information so that the expected benefits are greater than the expected costs' (2001) – as Sperber et al. now put it, that we are epistemically vigilant.

On this reading of the argument, its core idea is that we need to assume that humans are epistemically vigilant in order to account for the evolutionary stability of communication. Given that communicators frequently have an incentive to lie, non-vigilant recipients would end up forming many false testimonial beliefs, reducing their fitness until communication collapsed. Since communication appears to be evolutionarily stable, we can infer that recipients are vigilant.

1.2 *Vigilance as dispensable but adaptive*

The first version of the argument requires the strong assumption that dishonest testimony is sufficiently frequent to render a policy of non-vigilant acceptance sufficiently unreliable to produce a dramatic reduction in the fitness of receivers. Elsewhere, Sperber et al. suggest,

less starkly, that misinformation may be frequent enough to 'reduce, cancel, or even reverse' the potential gains to be had by accepting testimony (2010: 360). The possibility that dishonesty could reverse the gains provided by communication suggests the first reading of the argument, but the possibility that it could merely reduce or cancel them suggests an alternative reading, on which vigilance is not indispensable but is nevertheless adaptive. The core idea of this version of the argument would be that, since communicators inevitably produce some dishonest testimony, receivers benefit if they can filter out dishonest testimony, even if it would nevertheless still be adaptive simply to accept all incoming testimony.

On this reading, the argument no longer requires the strong assumption that dishonest testimony is sufficiently frequent to render non-vigilant testimonial belief formation highly unreliable, but still gives us reason to suppose that humans have evolved to be epistemically vigilant. And since, as we will see, there are both empirical and theoretical reasons to take dishonesty to be infrequent (though not, as Sperber et al. suggest (2010: 368), because recipients' vigilance prevents communicators from lying – see section 4), there is reason to favour the second version of the argument.

Note, however, that the second version depends on the additional assumption that the benefits of vigilance to the receiver outweigh its costs – this further fact follows if vigilance is indispensable to the stability of communication, but if vigilance need not be invoked to account for the stability of communication and is rather predicted because it would increase the benefits of communication to receivers, then it is required as a separate assumption that this benefit of vigilance is not outweighed by the (cognitive or other) costs of vigilance.

2. FORMS OF VIGILANCE

A range of forms of vigilance are possible. Considering various possible forms of vigilance allows me to specify two versions of the view which Sperber et al. might mean to defend, and to identify the most promising alternative view.

2.1 *Strongly, moderately and weakly effective vigilance*

When Sperber et al. (2010) argue that humans are epistemically vigilant, they might, at the most general level, be making either of two claims:

1. *Bare vigilance*: Recipients are vigilant in the sense that they monitor for (are on the lookout for) deception on the part of communicators, whether or not they often succeed in detecting it when it is present.
2. *Effective vigilance*: Recipients are vigilant in the sense that they monitor for deception on the part of communicators (they exercise bare vigilance), and this monitoring is effective, i.e. they often[2] succeed in detecting deception when it is present.

As I argue in section 3.2, the empirical studies cited by Sperber et al., since they provide little evidence of sensitivity to trustworthiness, primarily support the claim that we exercise bare vigilance. Their evolutionary argument, on the other hand, to the extent that it

2 As Sperber et al. (2010) point out, vigilance need not succeed in weeding out all dishonest communication. See below on quantitative dimensions of vigilance.

succeeds, supports the claim that we exercise effective vigilance, since it appeals to the benefits to receivers of sensitivity to trustworthiness.

These two lines of reasoning might seem to complement each other: the evolutionary argument shows that we must be effectively vigilant; the empirical research identifies mechanisms for bare vigilance; together, they suggest that the identified mechanisms are effective. But this appearance is misleading, for there is a large body of research on human deception detection which shows precisely that we are *not* effectively vigilant to any significant extent. And if we are not effectively vigilant, then obviously the evolutionary argument purporting to show that we are has gone wrong somewhere. Thus we are no longer entitled to conclude that the mechanisms for bare vigilance identified by Sperber et al. are likely to be effective.

Sperber et al. acknowledge the basic finding of deception detection research ('what it shows, in a nutshell, is that detecting deception on the basis of non-verbal behavioural signs is hard'), but argue that this does not spell trouble for vigilance: 'In order to gain a better grasp of the mechanisms for epistemic vigilance towards the source, what is most urgently needed is not more empirical work on lie detection or general judgments of trustworthiness, but research on how trust and mistrust are calibrated to the situation, the interlocutors and the topic of communication' (2010: 370–1). This move, however, only makes sense as long as we are assuming that vigilance must be effective to a significant extent and, as I will argue (section 3.1), work on lie detection strongly suggests precisely that trust and mistrust are *not* calibrated to the situation, the interlocutors and the topic of communication.

Given the basic distinction between bare and effective vigilance, there appear at first glance to be four possibilities that we need to consider:

1. Recipients exercise bare vigilance and therefore usually avoid being deceived.
2. Recipients exercise bare vigilance but are nevertheless often deceived.
3. Recipients do not exercise bare vigilance and therefore are often deceived.
4. Recipients do not exercise bare vigilance but nevertheless usually avoid being deceived.

Sperber et al. focus primarily on possibilities 1 and 3. Since it is plausible, on evolutionary grounds, that we are not too frequently deceived, they are right to rule out possibility 3, and the same sort of evolutionary considerations rule out possibility 2.

What of possibility 4? While we have good evolutionary grounds for assuming that we are not too frequently deceived, a variety of mechanisms for minimizing deception have been identified in both the deception literature and the animal signalling literature, and we should not, until we have canvassed these other mechanisms, be prepared to conclude that vigilance plays a major role in allowing humans to avoid deception. Indeed, though I will grant that we exercise a certain form of vigilance and that this vigilance makes a certain contribution to the reliability of testimonial belief formation, I will argue that this contribution is in fact minor – in an important sense, possibility 4 is closer to the truth than is possibility 1.

While possibility 4 is closer to the truth, we can get closer yet; the truth is somewhere between possibilities 1 and 4. In order to see how this might be the case, note that the 'therefore' in possibility 1 conceals an additional possibility, namely, that while recipients exercise bare vigilance and usually avoid being deceived, avoidance of deception is due primarily to some factor other than vigilance. The basic point here is that vigilance is not all-or-nothing but rather varies quantitatively along several dimensions. First, obviously,

we can ask: how successful is vigilance at detecting deception? The formulations given above already take this dimension into account. But there are at least two additional quantitative dimensions to vigilance:

- What are the relative contributions of vigilance and other factors to the reliability of communication?
- To what extent are receivers vigilant?

Focusing for the moment on the former question, we can distinguish the following possibilities.

1. *Strongly effective vigilance*: Recipients exercise bare vigilance, and, due primarily to this, they usually avoid being deceived.
   2–4 as above.
5. *Weakly effective vigilance*: Recipients exercise bare vigilance, and they usually avoid being deceived, but this is due primarily to some other factor.

Version 1 of the evolutionary argument supports possibility 1, strongly effective vigilance. Version 2, since it allows that vigilance need not be indispensable to reliability, supports an intermediate possibility:

6. *Moderately effective vigilance*: Recipients exercise bare vigilance, and they usually avoid being deceived; both vigilance and some other factor make significant contributions.

We have already ruled out possibilities 2 and 3 on basic evolutionary grounds. Sperber et al. provide convincing evidence that we are vigilant to some extent, which rules out possibility 4. But nevertheless possibility 4 is closer to the truth than either possibility 1 or possibility 6, for, while vigilance does play a role in enabling reliable testimonial belief formation (contra possibility 4), it plays only a minor role – in fact, vigilance is not strongly effective or even moderately effective but only weakly effective (possibility 5).

2.2 *Type 1 and type 2 vigilance*

In order to see how vigilance might play this sort of limited role, it is necessary to take the remaining quantitative dimension of vigilance into account: to what extent are we vigilant? The question has two parts. How often (when) are we vigilant? And how many resources do we devote to vigilance?

Sperber et al. take vigilance to be our default state – the mechanisms for vigilance are always on. Thus, though they do assign a role to what I will refer to as type 2 (reflective, conscious, deliberate) vigilance, they take vigilance to be primarily a matter of type 1 (automatic, unconscious, heuristic) processing, as far as vigilance with respect to honesty is concerned. While they do not explicitly draw the type 1/2 distinction (Frankish 2010), they remark in various places that monitoring is typically an unconscious, ongoing process, only occasionally giving rise to conscious, controlled processing, so they appear to have roughly this distinction in mind.

I will argue that, to the extent that we are vigilant with respect to honesty, vigilance is primarily a type 2 process which is engaged relatively rarely but which is highly reliable

when engaged. This sort of selectively triggered type 2 vigilance can make a modest contribution to the reliability of testimonial belief formation, even if it is not sufficient on its own to filter out very much deceptive communication. Though Sperber et al. do assign a limited role to selectively triggered type 2 vigilance, their view of this role differs somewhat from mine. They rely here on the argumentative theory of reasoning developed recently by Mercier and Sperber, according to which 'reasoning contributes to the effectiveness and reliability of communication by enabling communicators to argue for their claim and by enabling addressees to assess these arguments' (2011: 71–2). The basic claim is that interpretation of an utterance involves activation of relevant background beliefs, which may reveal that the communicated information is incoherent with the subject's existing beliefs. When this happens, a vigilant subject may engage in further coherence checking in order to determine whether to reject the communicated information or, instead, revise some of his other beliefs in order to accept it. I grant that such selectively triggered coherence checking might allow agents to filter out some dishonest communicated information, but point out that coherence checking will not be the whole story about type 2 vigilance, which can also be triggered by other factors, including the sort of behavioural cues discussed below (section 4.1). More importantly, I maintain, selectively triggered type 2 vigilance can make only a minor contribution to the reliability of testimonial belief formation, even if it is itself highly reliable, for, as Sperber remarks elsewhere (2001), coherence checking is cognitively costly and thus will be invoked only rarely, and this goes for type 2 vigilance in general. It is not clear whether Sperber et al.'s and Mercier and Sperber's view on coherence checking is consistent with this. They do refer to coherence checking as a mechanism that functions in addition to calibration of trust in the situation, so their view may be compatible with the point. On the other hand, their extensive discussion of coherence checking suggests that, as far as vigilance with respect to competence is concerned, they may intend to assign a more central role to type 2 vigilance. Since my argument focuses on vigilance with respect to honesty, I will not explore this further; as far as vigilance with respect to honesty is concerned, I differ from them primarily over the role of type 1 vigilance.

3. STRONGLY AND MODERATELY EFFECTIVE TYPE 1 VIGILANCE

In this section, I first review evidence that vigilance is neither strongly nor moderately effective. I then consider the positive evidence adduced by Sperber et al., arguing that it fails to support either strongly or moderately effective vigilance. In short: vigilance is at most weakly effective.

3.1 The evidence from deception detection research

Any mechanism for face-to-face deception detection will have two basic components: cues to deception provided by the communicator,[3] and a sensitivity to those cues on the part of the receiver.

3 The cues relevant in this context must be detectable by agents unaided, i.e. without the use of specialized equipment. The existence of internal differences between lying and honesty (e.g. changes in brain

3.1.1 Cues to deception

As the deception detection literature unambiguously shows, there are only a few, weak cues to deception. I have discussed research on cues elsewhere (Michaelian 2010); here, I briefly review the main findings, relying mainly on Vrij's recent authoritative survey (2008). Vrij argues that we have little theoretical reason to expect to be able to identify many cues strongly correlated with deception, since (1) the fact that someone lies need not change his behaviour, and (2) behaviours that accompany lying can also accompany truth-telling. And indeed, based on an extensive review of studies of cues to deception, Vrij concludes that there are only a small number of weak cues to deception.

Gaze aversion provides a vivid example. Folk wisdom has it that liars tend to avert their gaze: in a large-scale study carried out in 75 different countries and 43 different languages, a majority (64%) of people reported believing that 'liars look away' (Global Deception Research Team 2006). The belief that gaze aversion is associated with deception is intuitively plausible – lying is considered to be morally unacceptable, so liars should feel ashamed. Indeed, as Vrij 2008 points out, the belief is so plausible that it is regularly expressed in popular media and even police training manuals. However plausible the belief might be, gaze aversion turns out not to be a cue to deception. Of 46 studies reviewed by Vrij in which gaze aversion was investigated, only five found that liars avert their gaze more often than truth-tellers. Six found that truth-tellers avert their gaze more often than liars, and the remaining 35 studies found no relationship between honesty and gaze aversion. In their meta-analysis, DePaulo et al. (2003) find an effect size d of only .03 for gaze aversion. In general, one would expect liars to display many more behaviours associated with nervousness – hesitations, pauses, etc. – but these behaviours also turn out not to be genuine cues to deception. Similarly, we tend to think that lies are less consistent than truths, and that they contain more contradictions; this also turns out not to be the case. There are, however, some characteristic behavioural differences between liars and truth-tellers (though in many cases the effect size is small); for example, pitch of voice tends to differ, and certain movements are more characteristic of truth-tellers than liars. Similarly, lies tend to be shorter and less plausible than honest statements (Vrij 2008).

3.1.2 Sensitivity to cues

Since there are only a small number of genuine cues to deception, and since in many cases these are only weakly correlated with deception, it is unlikely that we can actually use cues to detect deception in real time, as strong or moderate vigilance would require. There are, however, some genuine cues to deception, even if these are weak, so we cannot rule the assumption out entirely – it is possible that, despite the difficulty of the task, agents manage to exploit the available cues to monitor effectively for deception. As the deception detection literature shows, however, this is not the case; deception detection accuracy

activity, measurable by fMRI (Wolpe et al. 2010)) or of external differences (e.g. changes in blood pressure and other physiological changes, measured in polygraph exams (National Research Council 2003)) which need not produce observable behavioural effects is irrelevant. What matter are perceptually detectable behavioural (verbal or non-verbal) cues, so I will consider only such cues in what follows.

has consistently been found to be barely above chance. Vrij (2008) finds an average accuracy rate of 54.25 per cent, and I will follow him in taking deception detection accuracy to be about 54 per cent.

Despite the consistency with which deception detection accuracy is found to be barely better than chance, one might attempt to dismiss the finding, insisting that the relevant experiments (a typical set-up requires subjects to provide deception judgements for videotaped statements by strangers) stack the deck against receivers, making it more difficult for them to detect deception than it normally is in ecological situations. This strategy is adopted by von Hippel and Trivers (2011), whose argument that self-deception evolved to facilitate other-deception, like Sperber et al.'s argument for vigilance, requires that sensitivity to deception is good. Von Hippel and Trivers admit that the literature seems to show that deception detection ability is poor, but suggest various ways in which it might underestimate agents' actual detection ability. For example, they point out that most studies are conducted using subjects who are strangers to each other, which eliminates the possibility of using cues that are specific to an individual and might be known to someone in an ongoing relationship with that individual. They then argue that '[i]f rates of deception detection are, in fact, substantially higher outside the laboratory than in it, we are led back to the notion of a co-evolutionary struggle between deceiver and deceived' (2011: 4).

If deception detection rates are substantially higher than the research apparently shows them to be, then we are indeed back to the 'evolutionary arms race' picture of communication. However, their dismissal of the relevant research is much too quick. They identify various ways in which typical deception detection studies depart from ecological situations and suggest that, if studies were more ecologically valid, detection rates would increase significantly, but they provide little evidence to support this. Moreover, they overlook the findings of existing studies in which their concerns are addressed. Similar points are emphasized by Vrij 2011 and Dunning 2011. Vrij points out that the claim that detection rates would go up in studies using subjects who are known to each other is contradicted by work on deception detection in close relationships (2011: 41). Dunning similarly points out that von Hippel and Trivers's claim that deception detection ability outside the laboratory is good is contradicted by the available evidence.

This raises the challenge of explaining why deception detection ability is poor. As Park and Levine (2001) emphasize, it is misleading to say simply that deception detection accuracy is slightly better than 50 per cent; in fact, deception detection accuracy varies as a function of the base rate of honesty, with deception detection accuracy increasing as the base rate of honesty increases (Levine et al., 1999). In order to account for this 'veracity effect', Park and Levine argue that subjects are in general truth-biased (Park and Levine 2001; Levine et al. 2006): they tend to judge that received messages are honest, regardless of actual message honesty.

This, in turn, raises the challenge of accounting for the existence of the truth bias. Here, we can point to the influence of false beliefs about cues to deception (there is only modest overlap between believed cues and genuine cues to deception: Vrij 2008) and, crucially, to the costs of monitoring for deception, which can be broken down as follows.

- Cognitive costs: as I argue in section 4, monitoring requires cognitive resources (over and above those required simply to interpret the utterance), giving agents an incentive not to monitor.

- Emotional costs: conversational norms prohibit certain behaviours (e.g. close questioning of the speaker) which can enable receivers to detect deception, entailing an emotional cost for agents who monitor for dishonesty, again giving agents an incentive not to monitor.
- Social costs: violation of these conversational norms may also result in social sanctions, giving agents an additional incentive not to monitor.

3.2 Sperber et al.'s evidence for effective vigilance

The evidence from deception detection research that receivers are rarely able to detect deception in face-to-face interaction is strong; and the positive evidence cited by Sperber et al. (on the formation of general judgements of trustworthiness, and on the development of children's attitudes towards testimony) does not provide significant support for strongly or moderately effective vigilance.

3.2.1 General judgements of trustworthiness

As Sperber et al. note, the trustworthiness of a given speaker will vary from situation to situation, implying that precise calibration of trust (allocation of trust to informants according to 'the topic, the audience, and the circumstances') would be beneficial. Such precise calibration, however, would be extremely costly, suggesting that we rely largely on general judgements of trustworthiness rather than precisely calibrated judgements. Since there is no 'failsafe way of calibrating one's trust in communicated information so as to weed out all and only the misinformation', they infer that there must have been 'ongoing selective pressure in favour of any available cost-effective means to at least approximate such sorting' (2010: 369).

The suggestion is that we must have some good-enough method of approximating highly accurate situation-by-situation evaluations of trustworthiness. The research to which Sperber et al. appeal in support of this suggestion, however, while it indeed confirms that we tend to form general judgements of trustworthiness using rapid, type 1 processes, tells us nothing (because it is not designed to tell us anything) about the accuracy of those judgements. For example, Willis and Todorov (2006) show that we form highly stable general judgements of trustworthiness in mere milliseconds, but this finding tells us nothing about the accuracy of those judgements. And if general judgements of trustworthiness are going to figure in an argument for effective vigilance, they must tend to be accurate.

Sperber et al. themselves note this problem, remarking that '[o]ne might wonder if such split-second judgements of trustworthiness have any basis at all' (2010: 370). And indeed one might: however tempting it is to suppose that formation of split-second general judgements of trustworthiness must be adaptive and that the judgements in question must therefore be fairly accurate, there is reason to suspect that they are not. Baby-faced people, for example, tend to be judged to be more trustworthy (Zebrowitz 1999), but there would appear to be no reason to take baby-facedness to correlate with actual trustworthiness. Sperber et al. argue in response that 'what this experiment suggests is that looking for signs of trustworthiness is one of the first things we do when we see a new face'. This is, of course, what the experiment shows, but, again, the finding that we form split-second general judgements of trustworthiness does not answer the question about the accuracy of those judgements.

Additionally, while Sperber et al. appeal to general judgements of trustworthiness as a means of approximating more accurate situation-by-situation evaluations, we might worry that formation of such general judgements is just a special case of our well-documented tendency to commit the fundamental attribution error, in which case we have an additional reason to expect general judgements not to be very accurate. Responding to this worry, Sperber et al. suggest that general judgements of trustworthiness might not be baseless, since different communicators might adopt different policies with respect to being honest. There are two problems with this strategy. First, there does not in fact seem to be much variation in the honesty policies that communicators adopt: while there appear to be a few 'prolific liars', most people are honest most of the time (Levine and Kim 2010; Serota et al. 2010), making it difficult to form judgements that track whatever slight differences there might be between honesty policies. Second, even if we suppose that there is significant variation among honesty policies, we have no evidence that our general judgements of trustworthiness track these differences.

3.2.2 The development of vigilance

The work on the development of children's attitudes towards testimony discussed by Sperber et al. concerns attitudes towards testimony where the trustworthiness of the communicator is already known or easily ascertained – in the relevant experiments, either the children are explicitly told that an informant is trustworthy or have an opportunity to develop a track-record for the informant, or the informant's testimony directly contradicts the children's own experience. But that children distrust testimony when they already know that the relevant testifier is untrustworthy is one thing; it is another thing for them to be able to reliably determine that a given testifier is untrustworthy (and so to distrust his testimony) when they do not already have that information. Due to this feature of the research, it provides very limited evidence about the reliability of children's vigilance: while it appears to show that, from a very young age, children are good at rejecting testimony where they have direct evidence about the trustworthiness of the speaker, it shows nothing (because it is not designed to show anything) about whether they reliably reject untrustworthy testimony where such direct evidence is unavailable. And given that only in a minority of our interactions do we have direct evidence about the trustworthiness of the speaker, the latter question is crucial.

Sperber et al. argue that young children are not gullible, blindly accepting whatever they are told, but rather allocate trust on the basis of available information about benevolence, appealing to a series of experiments by Mascaro and Sperber (2009), and competence, appealing to work by Clément and collaborators (2004). In Mascaro and Sperber's first experiment, 3-year-old children were told that a puppet was either 'kind' or 'mean'; it was found that they tended to prefer testimony from the kind puppet. Even granting that the children preferred testimony from the kind puppet because they took it to be more likely to be accurate, the experiment obviously does not show anything about their ability to detect deception, since they are given a reason to distrust the puppet ahead of time. In their second experiment, children were told that a puppet is a 'big liar' (that he 'always tells lies'); it was found that 3-year-old children tend nevertheless to accept the testimony of the puppet, whereas children 4 and older inferred that the opposite of what the puppet said was true. Again, this experiment tells us nothing about whether children are able to filter out dishonest testimony when they are not informed in advance that the testimony will be dishonest. In Mascaro and Sperber's final experiment, children were

told not that the puppet was a liar but that he was 'mean' and did not want them to find a sweet; it was found that older children are better at inferring that the mean puppet will attempt to mislead them about the location of the sweet. This experiment, like the others, tells us nothing about children's ability to detect deception where they do not already have reason to take it to be present, or even whether they attempt to do so.

Thus, while Mascaro and Sperber's research reveals much about the stages of children's developing understanding of deception and their ability to refrain from accepting the testimony of known deceptive informants, to the extent that epistemic vigilance is a matter of processes devoted to screening out incoming false information on the basis of available behavioural cues, the research does not provide any support for strongly or moderately effective vigilance.

The experiments by Clément et al. (2004) which Sperber et al. cite have analogous features, which means that they, too, do not provide significant support for effective vigilance.[4] In their first experiment, subjects (3-, 4- and 5-year-old children) were told that one puppet always gives the right answer, while the other always gives 'some strange answers'; they also had a chance to observe that the reliable puppet is reliable and that the unreliable puppet is unreliable. They found that 4- and 5-year-olds, if not the younger children, tended to prefer the testimony of the reliable puppet to that of the unreliable puppet and to prefer their own observations when these conflicted with the testimony of the reliable puppet. The set-up of their second experiment was similar, and they obtained similar results. Thus, while these two lines of research nicely complement each other, with one focusing on honesty and the other focusing on competence, neither demonstrates that children attempt to determine trustworthiness where it is previously unknown or where received testimony does not directly contradict observation; thus, even as far as bare vigilance is concerned, they concern only a limited form of vigilance. And thus neither demonstrates that children succeed in determining trustworthiness where it is previously unknown or where received testimony does not directly contradict observation; thus, as far as effective vigilance is concerned, they provide little support.

In a more recent article, Clément (2010) points out that relatively little work has been done on children's ability to filter communicated information when no information about the truth of the communicated information or about the speaker's reliability is available. In order to suggest that children are able to do this, he cites work showing that children rely to some extent on the age of the speaker (Jaswal and Neely 2006), on consensus (Corriveau et al. 2009), and on familiarity (Corriveau and Harris 2009) (in addition to the speaker's benevolence (Mascaro and Sperber 2009)). This research provides additional support for bare vigilance, but, like the work by Clément et al., it does not reveal anything about the reliability of children's judgements of trustworthiness in general or their judgements of honesty in particular. We might speculate that this work shows that children are sensitive to factors correlated with trustworthiness, which in turn suggests that their judgements of trustworthiness are somewhat reliable. But the research on deception detection in adults shows that we should not accept this sort of speculation. It is equally plausible, in the adult case, to argue that adults are sensitive to certain factors correlated with deception, which suggests that their deception judgements are fairly reliable, but, as we saw above (section 3.1), adults have only a slight sensitivity to deception.

4 Additionally, given that they focus on competence rather than honesty, they are only indirectly relevant.

Nor is there reason for special optimism with respect to children. While children's ability to lie has been investigated extensively, as has the ability of adults to detect children's lies, relatively little work has been done on children's own ability to detect lies. But the existing work suggests that children rely on cues similar to those that determine adults' deception judgements – in particular, that they rely on gaze aversion (Einav and Hood 2008). And adults' deception judgements are not sensitive in part because they rely on false cues, including gaze aversion, as we have seen. Thus it is unlikely that children's deception judgements are reliable. And so the research cited by Sperber et al. on the development of attitudes towards testimony, like the research on formation of general judgements of trustworthiness, provides no reason for us to accept strongly or moderately effective vigilance: these two lines of research show that, from an early age, we have a preference for not relying on untrustworthy communicators (the developmental work) and that we rely to some extent on type 1 processes to evaluate the trustworthiness of communicators (the work on general judgements of trustworthiness); but neither shows that we have a capacity for reliably detecting untrustworthiness.

3.3 Other sources of evidence

There are other sources of evidence which, while not cited by Sperber et al., could nonetheless be used to support their argument. In a recent study, Kogan et al. (2011) show that individuals who have a variation of a certain gene which is associated with higher levels of prosociality tend to be judged as being more prosocial (where this includes trustworthiness) by observers on the basis of brief (20 seconds) observations of their behaviour. One might argue that this finding reinforces Sperber et al.'s claim that we employ rapid type 1 processes enabling accurate judgements of trustworthiness: a genetic difference correlated with higher levels of trustworthiness gives rise to some observable differences which enable receivers to accurately judge trustworthiness.

But this would be premature at best, due to several features and limitations of the study. First, the study does nothing to undermine the finding that deception detection accuracy is poor, which by itself is sufficient to undermine Sperber et al.'s argument. Second, the targets who were judged to be more or less prosocial were not known to the experimenters to actually be more or less prosocial, so the study provides only indirect evidence that we are able to accurately judge prosociality on the basis of brief observations of behaviour. Moreover, since, as the authors themselves point out, prosociality is presumably determined by multiple factors, a single-gene approach like that taken in the Kogan et al. study can provide only limited evidence about sensitivity to prosociality. Finally, the study did not concern judgements of honesty on a particular occasion. Subjects were asked to rate target individuals for trustworthiness, rather than to evaluate their current statements for honesty. So, while it might provide some support for accurate judgements of trustworthiness of individuals, it provides no support for accurate judgements of trustworthiness of statements, since trustworthiness of an individual is only one among multiple factors determining trustworthiness on a particular occasion.

3.4 Against strongly and moderately effective vigilance

The first version of Sperber et al.'s argument is best understood as concluding that we employ type 1 processes enabling strongly effective vigilance – that is, that type 1 processes

output judgements that are highly reliable and so can play the major role in accounting for the reliability of testimonial belief formation. This view is not supported by the evidence cited by Sperber et al., and it is directly contradicted by the evidence from deception detection research. So we can conclude that the first version of the argument fails: we do not employ type 1 processes enabling strongly effective vigilance.

The second version of the argument is best understood as concluding that we employ type 1 processes enabling moderately effective vigilance – that is, that type 1 processes output judgements that are sufficiently reliable to play a significant role in accounting for the reliability of testimonial belief formation. This view, again, is not supported by the evidence cited by Sperber et al., nor is it consistent with the evidence from deception detection research, since this shows that our deception detection accuracy is barely better than chance. So we can likewise conclude that the second version of the argument fails: we do not employ type 1 processes enabling moderately effective vigilance.

4. WEAKLY EFFECTIVE TYPE 2 VIGILANCE

If vigilance is only weakly effective, there remains the question of how to account for our slight sensitivity to deception – is this the result of type 1 or type 2 processes? Once we have answered this question, it will remain to provide an alternative explanation of the evolutionary stability of communication.

4.1 Type 2 vigilance can explain deception detection accuracy

Given that selectively triggered type 2 vigilance is less costly than the alternative, always-on type 1 vigilance (see section 4.3), we should favour an explanation of our slight sensitivity to deception in terms of the former, if such an explanation is available. In this section, I draw on a recent proposal by Levine (2010) to argue that selectively triggered type 2 vigilance can indeed explain our slight sensitivity to deception.

Levine's core claim is that 'stable and slightly better than chance accuracy is a function of a few transparent liars' (2010: 44). It is usually supposed that our slightly better than chance accuracy (54%) should be explained in terms of 'leakage' – the idea that, because liars experience certain emotions not experienced by truth-tellers, and because they cannot fully control the expression of these emotions, lying creates cues that 'leak out', enabling receivers to detect lies (2010: 45). On this view, the detection accuracy rate is low because the leaked cues are imperfectly correlated with deception and because people also rely to some extent on false cues. However, as Levine argues, there are a number of findings that cannot be explained by this approach. If low accuracy rates were due mainly to use of false cues, training should improve accuracy, but training does not appear to improve accuracy significantly (2010: 47). Similarly, the leakage approach suggests that experience should improve lie-detection ability, with people such as police displaying superior ability, but experience, too, appears not to improve accuracy significantly (Levine 2010: 47). Finally, the leakage approach suggests that we should find significant variance in accuracy rates among receivers, but little variance is found. Levine's alternative explanation is that there are a few transparent liars, that is, liars whose behaviour makes it so easy to determine that they are lying that the receiver can easily do this, but that most liars are not transparent, a suggestion supported by the finding that variance

in communicator demeanour is larger than variance in receiver ability (Bond and Depaulo 2006).

The basic explanation is illustrated using a test-taking analogy (Levine 2010: 44–5). Suppose that a large number of students is given a series of tests consisting of 100 true–false questions, and that the average score on the test is around 55, with little variation among tests or students. One way to explain this pattern of scores is in terms of the abilities of the test-takers: most students learned around 10 per cent of the content to be tested; this allowed them to correctly determine the answers to 10/100 questions; guessing on the remaining 90 allows them to get about 45/90 right, for a result of 55/100 correct answers on the test. Given that there is little variance among scores, however, an alternative explanation emphasizing features of the test, rather than the test-takers, becomes plausible: it might be that 90 of the questions were so difficult that every student had to guess, regardless of his ability, while the remaining 10 questions were so easy that every student could get them right, again regardless of his ability. The suggestion is that, in an analogous manner, our slight sensitivity to deception should be explained in terms of variance in communicator performance, rather than in terms of receiver detection ability:

> most people can lie seamlessly without diagnostically useful leakage, but a few people tend to give themselves away and consequently are systematically detected by most observers. There are enough transparent senders to produce accuracy rates that statistically exceed chance level, but too few (under the conditions in most deception detection experiments) to allow for accuracy rates that exceed chance by much. Those few transparent senders are seen as transparent by almost everyone so there is much more variance in transparency than ability. Hence, accuracy findings tell us more about the sender than the detector. (Levine 2010: 49–51)[5]

If our slight sensitivity to dishonesty stems from the occasional transparency of liars, there is no need to invoke always-on type 1 monitoring to explain it; instead, the obvious cues given by transparent liars can trigger type 2 processes which enable accurate deception judgements on those occasions. This explanation requires that receivers have some degree of sensitivity to the cues provided by transparent liars, but the behaviour of transparent liars can make it apparent that they are lying, even if receivers do not employ a monitoring process dedicated to detecting cues: unusual behaviour by the communicator induces the receiver to consciously monitor him – and this type 2 monitoring, when engaged, permits formation of accurate deception judgements. Thus, if the 'few transparent liars' approach is right, our slightly better than chance accuracy may be explained in terms of type 2 rather than type 1 vigilance.

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 any significant resources to monitoring for attempts to deceive us, but rather (see section 4.2) concern ourselves with giving accurate information when we testify. Occasionally, however,

5 In section 3.1.2, I cited Park and Levine's probability model of deception detection accuracy (Park and Levine 2001), which appeals to receivers' truth bias to explain the veracity effect, in which accuracy is a function of the base rate of honesty. As Levine points out (2010: 52–3), the 'few transparent liars' model is consistent with the Park-Levine probability model: if truth-biased receivers were totally insensitive to deception, then at a .5 base rate of honesty, accuracy would be 50%; communicator transparency is invoked to explain the gap between this and the finding that accuracy is 54% at a .5 base rate of honesty.

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. Selectively triggered type 2 vigilance does not depend on ongoing type 1 vigilance. The idea is not that type 2 vigilance kicks in when the cues for which type 1 vigilance monitors cross a certain threshold but rather that the basic level of attention required to receive and interpret testimony is enough to reveal the behaviours that make type 2 vigilance appropriate.

4.2 *The base rate of deception is low*

Regardless of how our slight sensitivity to deception is to be explained, there remains the question of how to account for the reliability of communication given that we are largely insensitive to deception. Sperber et al. focus on the potential role of vigilance in ensuring the reliability of communication. But if the base rate of dishonesty is sufficiently low, then vigilance need not be invoked to explain the reliability of communication.

There are two points to be made here. First, there is empirical evidence that the base rate of deception is low (though admittedly this evidence by itself is too limited to be conclusive). Existing attempts to get a sense of the frequency of lying in non-laboratory settings all suggest that lying is a regular but infrequent occurrence. As noted above, while there is some evidence that there are a few prolific liars, most people appear to lie infrequently (Levine and Kim 2010; Serota et al. 2010; DePaulo et al. 1996; Levine et al. 2010). Second, given that we are largely insensitive to deception, the base rate must be low: otherwise, communication would collapse, but communication appears to be evolutionarily stable. In short, we can infer that the base rate of deception is low from the fact that receivers are largely insensitive to deception.

If the base rate of dishonesty is low, what prevents communicators from 'defecting'? On Sperber et al.'s picture, receiver vigilance compensates for communicator dishonesty – communicators defect but, due to receiver vigilance, communication is nevertheless reliable. On the alternative picture developed here, receiver vigilance cannot compensate for communicator dishonesty, but since dishonesty is relatively rare, this does not undermine the reliability of communication. But if vigilance is only weakly effective, what mechanism ensures that the base rate of dishonesty is low?

It is useful, in this context, to consider a point made by Sperber in response to a possible Reidian objection to his view. The Reidian might argue that, just as cooperation is an evolutionarily stable strategy in the iterated prisoners' dilemma, honesty is, for similar reasons, an evolutionarily stable strategy in the communication game. The suggestion is that, while it will often be in the communicator's short-term interests to lie (defect), it will normally be in his long-term interests to tell the truth (cooperate), since honesty is a prerequisite for developing a desirable reputation, which in turn affects the communicator's future ability to have receivers accept the information that he communicates. Responding to this objection, Sperber points out that there are a variety of situations in which the communicator's long-term interests will in fact be best served by dishonesty. Sometimes, a lie is more credible than the truth. Receivers do not care only about the truth – other factors, such as loyalty, are also relevant. And long-term effects do not always trump short-term effects – the

fact that a communicator lied once does not mean that he will lie again, and in many cases a communicator has information not available from other sources, leaving receivers with no choice but to depend on his information (Sperber 2001).

While there are indeed situations in which communicators' long-term interests are best served by dishonesty, Sperber overestimates their impact. Again, this must be the case: given receivers' poor deception detection ability, it cannot be the case that speakers' interests result in frequent deception. Sperber's discussion already hints at part of the mechanism ensuring communicator honesty, when he grants that the need to secure a reputation for honesty matters (social costs of deception), though he downplays its importance. And there are additional costs of deception not discussed by Sperber: emotional costs, related to the emotions accompanying deception, emphasized by the leakage approach;[6] and, crucially, cognitive costs. In the remainder of this section, I argue that these costs give rise to an honesty bias, which ensures that the base rate of deception is low.

4.2.1 Handicaps, indices and deterrents

It is potentially useful, when searching for explanations for the reliability of human communication, to draw on ideas from animal signalling research, but caution is required when doing so – as Sperber et al. point out, there are disanalogies between animal signalling and human communication, in virtue of which explanations of the reliability of animal signalling cannot automatically be transferred to human communication. But neither should we adopt an overly restrictive view of the available explanations, and risk concluding prematurely that the reliability of human communication cannot be explained by mechanisms similar to those underwriting the reliability of animal signalling.

As a number of animal signalling researchers have pointed out (see Cronk 2005; Grose 2011; Számadó 2011), researchers in other areas have mistakenly tended to assume that the handicap principle (Zahavi and Zahavi 1997) tells the whole story about the reliability of animal signalling; they have, moreover, tended to misinterpret the handicap principle. According to the handicap principle, signals are guaranteed to be honest because they are costly and therefore hard to fake. As noted in section 1, most human communication, unlike a peacock's tail, is not costly to produce, so the handicap principle appears not to get a grip in the human case. But as the researchers cited above have emphasized, it is important to note that what matters here is differential cost, rather than absolute cost: an impressive tail is more difficult for a low-quality male to produce than for a high-quality male. Even if we bear this point in mind, however, the handicap principle still does not get a grip in the case of human communication. In signalling governed by the handicap principle, greater costs are borne by honest signallers (the peacock with an impressive tail pays the costs of the tail), but it is typically not more costly to produce an honest utterance than it is to produce a dishonest utterance.[7]

This is where it becomes important not to overlook alternatives to the handicap principle (Smith and Harper 1995). Scott-Phillips (2008) emphasizes that, even within the

6 While these emotions might not result in significant leaked cues, they may nevertheless be experienced by the agent.

7 There may be some cases where the handicap principle applies to human communication: certain complex forms of speech could be used to convey information about the quality of the speaker (Locke 2008), and polite speech might be a handicap (van Rooy 2003). But these are clearly special cases.

domain of animal signalling, the handicap principle has only limited application. Consider e.g. sparrows, which use dominance badges; these appear not to entail significant costs but nevertheless are reliably correlated with fighting ability (Scott-Phillips 2008). He groups the other mechanisms that have been identified for ensuring reliability as follows: (1) indices in which meaning is tied to form (Smith and Harper 2003; Searcy and Nowicki 2005 consider the example of fundamental frequency of vocalization, where frequency is constrained by vocal cord length, which is correlated with body size); (2) deterrents in which costs are payed by dishonest signallers rather than honest signallers (as in handicap cases).

Human utterances are obviously not indices, but Scott-Phillips argues that deterrents, in the form of punishment for dishonesty (refusal to engage in further interactions) can account for the reliability of human communication.[8] Drawing on a model developed by Lachmann et al. 2001, he argues that '[s]ufficient conditions for cost-free signalling in which reliability is ensured through deterrents are that signals be verified with relative ease and that costs be incurred when unreliable signalling is revealed' and that these conditions are satisfied in the human case. The condition that costs are incurred when dishonest (unreliable) signalling is revealed indeed seems safe. Though, as Sperber points out (2001), there will be cases in which receivers effectively have no choice other than continuing to interact with a communicator despite knowing that he has lied, as well as cases in which receivers elect to continue to interact with a communicator despite knowing that he has lied, it seems likely that such cases are exceptional. Once one develops a reputation for lying, it will become difficult for one to have one's testimony taken up by receivers. The condition that signals be verified with relative ease will likewise often be satisfied. Though dishonesty is rarely detected at the time of the utterance, it can often be detected after the fact, by means including physical evidence and third-party testimony (Park et al. 2002; Michaelian 2010). (The capacity for such after-the-fact deception detection does not amount to a form of vigilance, in the relevant sense, but only to a normal sensitivity to evidence: as we saw in section 2, what is at issue here is our capacity for real-time detection of deception.)

4.2.2 The costs of deception

There will, however, also be many cases in which testimony cannot easily be verified, so one might worry that an explanation of reliability in terms of deterrence ultimately does not work. The fact that source memory is generally poor also suggests that there will be many cases in which punishment will not be incurred even when testimony is discovered to be false, since the receiver might no longer remember which communicator provided the testimony. However, the deterrents relevant to human communication include not only the sort of *social costs* – punishment of dishonest communication – considered by Scott-Phillips (and Sperber, in his discussion of the Reidian objection), but also other costs of dishonesty, imposed not by the receivers with whom the communicator interacts but rather by the communicator's own make-up. The list of relevant deterrents should be expanded to include not only punishment (social costs) but also emotional and cognitive costs (in parallel to the costs of monitoring discussed in section 3.1.2).

8 A similar argument about deterrence is given by Giardini 2011, focusing on the case of gossip.

Lying will, in normally socialized agents, have an *emotional cost*, since they have internalized norms that forbid lying, except under special circumstances. Violation of such internalized norms constitutes a disincentive to lying. Though they do not play the major role, I note that emotional costs based on violation of social norms against lying can be invoked in the sort of evolutionary explanation developed here, since there is no requirement that, in such an explanation, the low base rate of deception be explained entirely in terms of built-in factors internal to communicators – socially imposed deterrents can play a role, as can internal factors derived from social pressures.

More importantly, lying entails *cognitive costs*. Vrij et al. (2011: 28–9) argue that, for a number of reasons, deception is cognitively more costly than honesty:

- Formulating a lie may be more cognitively demanding, since coherence cannot be taken for granted.
- Liars generally devote more resources to monitoring and controlling their own behaviour, since they are less likely to take their credibility for granted.
- Liars generally devote more resources to monitoring receivers, for the same reason.
- Liars may have to remind themselves to role-play.
- Liars have to actively suppress the truth (which tends to be activated automatically).
- Liars have to deliberately activate the lie (the lie tends not to be activated automatically).

This view receives support from the work of Vrij's group showing that lies become easier to detect when the liar's cognitive load is increased (Vrij et al. 2006, 2008, 2011), and there are additional sources of support for the view. Verschuere et al. 2010 point out that imaging studies (Christ et al. 2009; Spence and Kaylor-Hughes 2008) support the view that honesty is the default, with no area of the brain more active for honesty than for deception. Similarly, drawing on work on deception detection, developmental work and imaging studies (Johnson 2004; Mohamed et al. 2006), Gombos 2006 builds a case that lying normally involves the use of effortful executive processes.

Thus it appears that deterrence can explain the low base rate of deception in human communication; crucially, the relevant deterrents do not depend on significant sensitivity to deception on the part of receivers.

4.3 The costs of monitoring

Given that the base rate of deception is low (and that significant receiver sensitivity to deception is not required to bring this about), a policy of default acceptance (the observed truth bias) is adaptive, as Levine 2010, among others (O'Sullivan 2003), has argued.[9] Since it will lead the agent to accept some inaccurate information, such a policy will result in a slight loss in reliability, relative to what could be attained by effective vigilance, assuming that such vigilance is possible in principle. But this loss in reliability will be offset by an increase in power (in the sense of Goldman 1992), resulting from the increased amount of accurate information accepted by the agent.

9 It is possible that, while a strong truth bias was adaptive in an ancestral environment in which we lived in small groups, permitting good after-the-fact deception detection, it is no longer adaptive in our actual conditions, where we interact with much larger numbers of people and have many more one-time interactions.

This view is reinforced when we consider the relative costs of feasible policies for response to testimony: a policy of default acceptance is cheaper than a policy of default monitoring, and the loss in reliability that it entails can moreover be minimized if the agent employs monitoring in a selective way. The alternatives here are a policy of default acceptance with contextually triggered type 2 monitoring and a policy of default type 1 monitoring. While type 1 monitoring is cheaper than type 2 monitoring, a policy requiring always-on type 1 monitoring will be more costly than a policy which requires only sparing use of type 2 monitoring. Additionally, selectively triggered monitoring avoids the emotional and social costs mentioned in section 3.1.2. We have seen that such selective monitoring can account for our slight sensitivity to deception (section 4.1). Thus it appears that we should explain our slight sensitivity to deception in terms of such monitoring.

In short, I argue, rather than strongly or moderately effective type 1 vigilance, only weakly effective type 2 vigilance plays an important role in accounting for the reliability of communication. This role is minor relative to that of the honesty bias, but it still allows us to account, in evolutionary terms, for the fact that we exercise a form of vigilance, since, while the benefits of contextually triggered type 2 monitoring are relatively small, so are the costs, making such a form of vigilance adaptive. The overall picture that emerges is one on which the reliability, and hence the evolutionary stability, of communication is ensured largely by the honesty bias, with type 2 monitoring playing a limited role when the truth bias is overcome by the cues provided by transparent liars.[10]

REFERENCES

Bond, Charles F., and Depaulo, Bella M. 2006. 'Accuracy of Deception Judgments.' *Personality and Social Psychology Review*, 10(3): 214–34.

Christ, Shawn E., Van Essen, David C., Watson, Jason M., Brubaker, Lindsay E., and McDermott, Kathleen B. 2009. 'The Contributions of Prefrontal Cortex and Executive Control to Deception: Evidence from Activation Likelihood Estimate Meta-Analyses.' *Cerebral Cortex*, 19(7): 1557–66.

Clément, Fabrice. 2010. 'To Trust or Not to Trust? Children's Social Epistemology.' *Review of Philosophy and Psychology*, 1(4): 531–49.

——, Koenig, Melissa, and Harris, Paul. 2004. 'The Ontogenesis of Trust.' *Mind and Language*, 19(4): 360–79.

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

Corriveau, Kathleen, and Harris, Paul L. 2009. 'Choosing your Informant: Weighing Familiarity and Recent Accuracy.' *Developmental Science*, 12(3): 426–37.

——, Fusaro, Maria, and Harris, Paul L. 2009. 'Going with the Flow: Preschoolers Prefer Nondissenters as Informants.' *Psychological Science*, 20(3): 372–7.

Cronk, Lee. 2005. 'The Application of Animal Signaling Theory to Human Phenomena: Some Thoughts and Clarifications.' *Social Science Information*, 44(4): 603–20.

Dawkins, R., and Krebs, J. R. 1978. 'Animal Signals: Information or Manipulation?' In J. R. Krebs and R. N. Davies (eds), *Behavioural Ecology: An Evolutionary Approach*, pp. 282–309. Oxford: Blackwell Scientific.

10 Thanks for comments to Alvin Goldman, and anonymous reviewer, and participants in seminars at Lund (organized by Frank Zenker and Bengt Hansson) and Bilkent (organized by Sandrine Berges).

DePaulo, Bella M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., and Epstein, J. A. 1996. 'Lying in Everyday Life.' *Journal of Personality and Social Psychology*, 70(5): 979–5.

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

Dunning, David. 2011. 'Get Thee to a Laboratory.' *Behavioral and Brain Sciences*, 34(1): 18–19.

Einav, Shiri, and Hood, Bruce M. 2008. 'Tell-Tale Eyes: Children's Attribution of Gaze Aversion as a Lying Cue.' *Developmental Psychology*, 44(6): 1655–67.

Frankish, Keith. 2010. 'Dual-Process and Dual-System Theories of Reasoning.' *Philosophy Compass*, 5(10): 914–26.

Giardini, Francesca. 2011. 'Deterrence and Transmission as Mechanisms Ensuring Reliability of Gossip.' *Cognitive Processing* (Oct.): 1–11.

Global Deception Research Team. 2006. 'A World of Lies.' *Journal of Cross-Cultural Psychology*, 37 (1): 60–74.

Goldman, A. 1992. *Liaisons*. Cambridge, MA: MIT Press.

Gombos, Victor A. 2006. 'The Cognition of Deception: The Role of Executive Processes in Producing Lies.' *Genetic, Social, and General Psychology Monographs*, 132(3): 197–214.

Grose, Jonathan. 2011. 'Modelling and the Fall and Rise of the Handicap Principle.' *Biology and Philosophy*, 26(5): 677–96.

Jaswal, Vikram K., and Neely, Leslie A. 2006. 'Adults Don't Always Know Best.' *Psychological Science*, 17(9): 757–8.

Johnson, R. 2004. 'The Contribution of Executive Processes to Deceptive Responding.' *Neuropsychologia*, 42(7): 878–901.

Kogan, Aleksandr, Saslow, Laura R., Impett, Emily A., Oveis, Christopher, Keltner, Dacher, and Saturn, Sarina Rodrigues. 2011. 'Thin-Slicing Study of the Oxytocin Receptor (OXTR) Gene and the Evaluation and Expression of the Prosocial Disposition.' *Proceedings of the National Academy of Sciences of the United States of America* (Nov.).

Krebs, J. R., and Dawkins, R. 1984. 'Animal Signals: Mind-Reading and Manipulation.' In J. R. Krebs and N. B. Davies (eds), *Behavioural Ecology: An Evolutionary Approach*, pp. 380–402. Sunderland, MA: Sinauer Associates.

Lachmann, Michael, Számadó, Szabolcs, and Bergstrom, Carl T. 2001. 'Cost and Conflict in Animal Signals and Human Language.' *Proceedings of the National Academy of Sciences of the United States of America*, 98(23): 13189–94 (Nov.).

Levine, Timothy R. 2010. 'A Few Transparent Liars: Explaining 54% Accuracy in Deception Detection Experiments.' In C. T. Salmon (ed.), *Communication Yearbook 34*, pp. 41–61. New York: Routledge.

——, and Kim, Rachel K. 2010. 'Some Considerations for a New Theory of Deceptive Communication.' In M. Knapp and M. McGlone (eds), *The Interplay of Truth and Deception*, pp. 16–34. New York: Routledge.

——, and Hamel, Lauren M. 2010. 'People Lie for a Reason: Three Experiments Documenting the Principle of Veracity.' *Communication Research Reports*, 27(4): 271–85.

——, Park, Hee S., and Hughes, Mikayla. 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): 243–60.

——, Park, Hee S., and McCornack, Steven A. 1999. 'Accuracy in Detecting Truths and Lies: Documenting the Veracity Effect.' *Communication Monographs*, 66(2): 125–44.

Locke, J. 2008. 'Cost and Complexity: Selection for Speech and Language.' *Journal of Theoretical Biology*, 251(4): 640–52.

McKay, Ryan T., and Dennett, Daniel C. 2009. 'The Evolution of Misbelief.' *Behavioral and Brain Sciences*, 32(6): 493–510.

Mascaro, Olivier, and Sperber, Dan. 2009. 'The Moral, Epistemic, and Mindreading Components of Children's Vigilance towards Deception.' *Cognition*, 112(3): 367–80.

Mercier, Hugo, and Sperber, Dan. 2011. 'Why do Humans Reason? Arguments for an Argumentative Theory.' *Behavioral and Brain Sciences*, 34(2): 57–74.

Michaelian, Kourken. 2010. 'In Defence of Gullibility: The Epistemology of Testimony and the Psychology of Deception Detection.' *Synthese*, 176(3): 399–427.

Mohamed, Feroze B., Faro, Scott H., Gordon, Nathan J., Platek, Steven M., Ahmad, Harris, and Williams, J. Michael. 2006. 'Brain Mapping of Deception and Truth Telling about an Ecologically Valid Situation: Functional MR Imaging and Polygraph Investigation Initial Experience 1.' *Radiology*, 238(2): 679–88.

National Research Council. 2003. *The Polygraph and Lie Detection*. Washington, DC: National Academies Press.

O'Sullivan, M. 2003. 'The Fundamental Attribution Error in Detecting Deception: The Boy-Who-Cried-Wolf Effect.' *Personality and Social Psychology Bulletin*, 29: 1316–27.

Park, Hee S., and Levine, Timothy. 2001. 'A Probability Model of Accuracy in Deception Detection Experiments.' *Communication Monographs*, 68(2): 201–10.

——, McCornack, Steven, Morrison, Kelly, and Ferrara, Merissa. 2002. 'How People Really Detect Lies.' *Communication Monographs*, 69(2): 144–57.

Scott-Phillips, Thomas C. 2008. 'On the Correct Application of Animal Signalling Theory to Human Communication.' In A. D. M. Smith, K. Smith, and R. Ferrer i Cancho (eds), Proceedings of the 7th International Conference on the Evolution of Language, pp. 275–82. World Scientific.

Searcy, W. R., and Nowicki, S. 2005. *The Evolution of Animal Communication: Reliability and Deception in Signalling Systems*. Princeton: Princeton University Press.

Serota, Kim B., Levine, Timothy R., and Boster, Franklin J. 2010. 'The Prevalence of Lying in America: Three Studies of Self-Reported Lies.' *Human Communication Research*, 36(1): 2–25.

Smith, M. J., and Harper, D. G C. 1995. 'Animal Signals: Models and Terminology.' *Journal of Theoretical Biology* (Dec.): 305–11.

—— and ——. 2003. *Animal Signals*. Oxford: Oxford University Press.

Smith, W. J. 1977. *The Behaviour of Communicating: An Ethological Approach*. Cambridge, MA: Harvard University Press.

Spence, Sean A., and Kaylor-Hughes, Catherine J. 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: 401–13.

——, Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., and Wilson, D. 2010. 'Epistemic Vigilance.' *Mind and Language*, 25(4): 359–93.

Stich, Stephen P. 1990. *The Fragmentation of Reason*. Cambridge, MA: MIT Press.

Számadó, Szabolcs. 2011. 'The Cost of Honesty and the Fallacy of the Handicap Principle.' *Animal Behaviour*, 81(1): 3–10.

van Rooy, Robert. 2003. 'Being Polite is a Handicap: Towards a Game Theoretical Analysis of Polite Linguistic Behavior.' *Proceedings of the 9th Conference on Theoretical Aspects of Rationality and Knowledge, TARK '03*, pp. 45–58. New York: ACM.

Verschuere, Bruno, Spruyt, Adriaan, Meijer, Ewout H., and Otgaar, Henry. 2010. 'The Ease of Lying.' *Consciousness and Cognition* (Nov.).

von Hippel, William, and Trivers, Robert. 2011. 'The Evolution and Psychology of Self-Deception.' *Behavioral and Brain Sciences*, 34(1): 1–16.

Vrij, Aldert. 2008. *Detecting Lies and Deceit: Pitfalls and Opportunities*. 2nd edn. Chichester: Wiley.

——. 2011. 'Self-Deception, Lying, and the Ability to Deceive.' *Behavioral and Brain Sciences*, 34(1): 40–1.

——, Fisher, Ronald, Mann, Samantha, and Leal, Sharon. 2006. 'Detecting Deception by Manipulating Cognitive Load.' *Trends in Cognitive Sciences*, 10(4): 141–2.

——, Granhag, Pär A., Mann, Samantha, and Leal, Sharon. 2011. 'Outsmarting the Liars: Toward a Cognitive Lie Detection Approach.' *Current Directions in Psychological Science*, 20(1): 28–32.

——, Mann, Samantha, Fisher, Ronald, Leal, Sharon, Milne, Rebecca, and Bull, Ray. 2008. 'Increasing Cognitive Load to Facilitate Lie Detection: The Benefit of Recalling an Event in Reverse Order.' *Law and Human Behavior*, 32(3): 253–65.

Willis, Janine, and Todorov, Alexander. 2006. 'First Impressions.' *Psychological Science*, 17(7): 592–8.

Wolpe, Paul R., Foster, Kenneth R., and Langleben, Daniel D. 2010. 'Emerging Neurotechnologies for Lie-Detection: Promises and Perils.' *American Journal of Bioethics*, 10(10): 40–8.

Zahavi, A., and Zahavi, A. 1997. *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.
Zebrowitz, L. A. 1999. *Reading Faces: Window to the Soul?* Boulder, CO: Westview.

KOURKEN MICHAELIAN is assistant professor in the philosophy department at Bilkent University (Ankara). His current research aims at developing empirically-grounded approaches to philosophical issues around memory and testimony, often focusing on the role of metacognitive monitoring and control processes.