Please quote the published version
Grodniewicz, J. P. (*forthcoming*) Effective filtering: Language comprehension and testimonial entitlement. *The Philosophical Quarterly*

Effective filtering: Language comprehension and testimonial entitlement

J. P. Grodniewicz
(j.grodniewicz@gmail.com)

Abstract: It is often suggested that we are equipped with a set of cognitive tools that help us to filter out unreliable testimony. But are these tools effective? I answer this question in two steps. Firstly, I argue that they are not real-time effective. The process of filtering, which takes place simultaneously with or right after language comprehension, does not prevent a particular hearer on a particular occasion from forming beliefs based on false testimony. Secondly, I argue that they are long-term effective. Some hearers sometimes detect false testimony, which increases speakers' incentive for honesty, and stabilizes the practice of human communication in which deception is risky and costly. In short, filtering prevents us from forming a large number of beliefs based on false testimony not by turning each of us into a high-functioning polygraph but by turning the social environment of human communication into one in which such polygraphs are not required. Finally, I argue that these considerations support strong anti-reductionism about testimonial entitlement.

Keywords: testimony; filtering; epistemic vigilance; anti-reductionism; deception detection

1. Introduction

A widespread opinion in the philosophical debate about testimony is that we are not helplessly gullible. We engage in some form of assessment of the reliability of testimony or, at least, we are sensitive to signs of its unreliability. This assumption plays a particularly important role in reductionist theories of testimonial entitlement, according to which, the warrant of testimony-based beliefs depends on hearers' having some reductive (i.e. based not on testimony itself but, for example, perception, memory, or inductive inference) reasons for trusting their testifier (e.g. Hume 1975; Adler 1994; Audi 1997, 2006; Fricker 1987, 1994, 1995, 2006; Lipton 1998; Lyons 1997). As stated in Elizabeth Fricker's classic paper:

> … the hearer should be discriminating in her attitude to the speaker, in that she should be continually evaluating him for trustworthiness throughout their exchange, in the light of the evidence, or cues, available to her. This will be partly a matter of her being disposed to deploy background knowledge which is relevant, partly a matter of her monitoring the speaker for any tell-tale signs revealing likely untrustworthiness. (Fricker 1994: 149-50)

On this view, monitoring (or filtering)[1] is a way of collecting evidence of speakers' trustworthiness.

The alternative to reductionism is anti-reductionism, according to which testimony-based beliefs can be warranted without hearers' possessing any reductive reasons (e.g. Reid 1970; Burge 1993, 1997; Coady 1992; Goldberg 2006, 2007, 2010, 2014; Goldman 1999; Graham 2006, 2010, 2015; Greco 2015; Simion 2020). Even though reductionists tend to accuse anti-reductionists of painting hearers as helplessly gullible,[2] filtering can be incorporated into anti-reductionist accounts of testimonial entitlement. For example, Sanford Goldberg (2007) offers the following formulation of anti-reductionism:

> A hearer H is epistemically justified in accepting… another's testimony on occasion O so long as (i) there are no undefeated good (doxastic, factual, or normative) reasons not to accept the testimony, and (ii) on O H's acceptance was the outcome of a process that exhibited a 'counterfactual sensitivity' to the presence of defeaters (which, given (i), turns up no such defeaters on O). (2007: 168)

Here, filtering takes the form of 'counterfactual sensitivity.' Goldberg (2007: 166) highlights that it is sufficient for hearers to *be on the lookout* for defeaters (as opposed to actively *looking for defeaters*) to avoid the charge of gullibility and that this condition is fully compatible with anti-reductionism.[3]

This paper takes up the question: 'Is filtering effective?' More precisely, is the following claim true?

> **Effective filtering (EF)**: Filtering reliably[4] prevents hearers from forming beliefs based on false testimony.

---

[1] Throughout the text I will use *monitoring* and *filtering* interchangeably. I will also use the noun 'filter' to refer to a set of mechanisms, processes, or any cognitive tools that allow us to perform filtering.
[2] Cf. Fricker (1994: 154, 1995: 404).
[3] Authors carving up a middle ground between reductionism and anti-reductionism also suggest that we routinely monitor received testimony (see, e.g. Lackey 2006).
[4] For the purpose of this discussion, it is sufficient to define reliability as requiring a high ratio of successes to failures. I will not attempt to specify precisely what 'high' means. Roughly, a high ratio is quite a lot higher than 0.5 but lower than 1.0. I briefly return to this issue at the end of Section 4.

My answer consists of two steps. In the first step (Section 2), I argue that filtering is not real-time effective, i.e. the following claim is false:

> **Real-time effective filtering (RTEF)**: Filtering reliably prevents hearers from forming beliefs based on false testimony because the process of filtering which takes place simultaneously with, or right after linguistic comprehension, reliably prevents a particular hearer on a particular occasion from forming beliefs based on a particular instance of false testimony.

To establish that RTEF is false, I argue, first, that Michaelian (2010, 2012, 2013) and Shieber's (2012, 2015) arguments against real-time effectiveness of filtering are immune to recent critique by Fricker (2016). Second, I outline and briefly defend an account of language comprehension, on which filtering *cannot* be real-time effective.

Even if filtering is not real-time effective, it does not mean that it is not effective at all. In Section 3, I argue that filtering is long-term effective:

> **Long-term effective filtering (LTEF):** Filtering reliably prevents hearers from forming beliefs based on false testimony because the fact that some hearers sometimes detect false testimony increases speakers' incentive for honesty and stabilizes the practice of human communication in which deception is risky, costly, and thus not very prevalent.

In short, filtering prevents us from forming (a large number of) beliefs based on false testimony (i.e. EF is true), not by allowing us to detect particular instances of false testimony (i.e. RTEF is false), but by decreasing the amount of false testimony which we encounter in our social environment (i.e. LTEF is true). My arguments for the latter claim draw from the work of Dan Sperber (Sperber et al. 2010; Sperber 2013) and Timothy Levine (2014, 2019b). Finally (Section 4), I argue that all these considerations support a version of anti-reductionism about testimonial entitlement:

> **Strong social anti-reductionism (SSAR):** Hearers are prima facie entitled to form beliefs based on what speakers assert because filtering is long-term effective.

2. Real-time effectiveness of filtering

2.1 Filtering is not real-time effective

Let us first examine whether filtering is real-time effective, i.e. whether EF is true because RTEF is true. Two authors who recently devoted a lot of attention to the real-time effectiveness of filtering are Kourken Michaelian (2010, 2012, 2013) and Joseph Shieber (2012, 2015). Both of them proceed from a compelling assumption: as long as we are interested in how filtering works and how impermeable it is, we cannot limit ourselves to theoretical considerations. Instead, we should consult relevant empirical research, mainly from the field of social psychology.

Michaelian's (2010) paper is a polemic with Fricker's reductionist theory of testimonial justification and knowledge (Fricker 1987 and onward). Michaelian breaks this theory into two main components: (i) the claim that the reductive account of testimonial justification (reduction of testimonial justification to receiver's possession of non-testimonial reasons) is necessary, and (ii) the claim that it is available. The question regarding the effectiveness of filtering plays a crucial role in assessing the second claim. According to Fricker, the process of formation of testimonial beliefs makes use of receivers' beliefs about the honesty and competence of the testifiers. These beliefs, in turn, are acquired by means of — plausibly *unconscious* (Fricker 1994: 150) — monitoring for signs of dishonesty and incompetence. What does this monitoring require? Fricker reassures us that it *does not* require '…an extensive piece of MI5-type "vetting" of any speaker before… [the hearer] may accept anything… [the speaker] says as true' (1994: 154). Instead, '[e]xpert dissimulators among us being few, the insincerity of an utterance is very frequently betrayed in the speaker's manner, and so is susceptible of detection' (Fricker 1994: 150).

Michaelian argues that even though Fricker might be right that we frequently and casually monitor for competence and insincerity, there are no good reasons to assume that this monitoring is effective. He focuses on detection of a speaker's dishonesty and presents a line-up of empirical studies which suggest that our accuracy rates in deception detection are 'only slightly better than fifty-fifty' (Levine et al. 1999: 126; cf. Bond and DePaulo 2006; Levine 2019a).[5] There may be many reasons for such underwhelming results, but Michaelian suggests that it is neither that receivers do not monitor for cues to deception nor that there are no cues to deception to monitor for in the first place. Rather, the task of monitoring for cues to deception is very difficult: receivers plausibly do

[5] The slightly better than chance accuracy (around 54%) might result from *a few transparent liars* effect (Levine 2010). According to Levine's metaphor, deception detection is similar to taking a test where you know answers to approx. 10% of the easiest questions and guess all the rest.

not monitor for all the cues to deception and the ones that they monitor for are subtle and vary significantly across agents and contexts (Vrij 2004; Feeley and Young 2000).

Shieber (2012, 2015) raises similar worries. The literature in psychology suggests neither that there is a unified set of cues to deception nor that we are particularly good at spotting them. Moreover, the accuracy rates of professionals whose work requires sensitivity to deception, such as police officers, are not significantly better than these of laypeople (Kraut and Poe 1980; DePaulo and Pfeifer 1986; Kohnken 1987). We are bad at deception detection by default, and there is little room for improving this skill through training.[6]

Fricker (2016) strongly resists the pessimism. What we have to take into account while assessing the effectiveness of filtering — she suggests — is that our communication takes place in different *testimonial environments* (T-environments). T-environments are individuated based on 'what frequency and what manner of false testimony... [the receiver] might easily encounter' (Fricker 2016: 96). For a filter to be effective is for it to be effective in a given T-environment; the same 'narrowly' construed belief forming method might turn out to be reliable in one environment (e.g. one's everyday T-environment full of family members, friends, and acquaintances), but not another (e.g. a T-environment full of habitual and expert liars). If the empirical studies were supposed to help us assess the effectiveness of filtering, environments created in the experimental settings would have to be sufficiently similar to the everyday T-environments of participants. But this is not the case in the available empirical literature. Most importantly, however, even if there were reasons to assume that we do not filter effectively, there are no reasons to assume that we cannot *learn* to filter effectively. According to Fricker, the jury is out on effectiveness of filtering until we demonstrate that '[h]umans are *constitutionally incapable* of learning to respond discriminatingly to testimony.' (2016: 103, *emphasis mine*).

Fricker concludes that data cited by Michaelian and Shieber is insufficient to vindicate the claim that filtering is not effective. This comes as no surprise, I think, given the requirements she imposes makes it virtually impossible to vindicate this claim using empirical methods. Fricker does not explain in detail how finely we should individuate T-environments, however, given that T-environments are individuated by appeal to frequency and manner of false testimony one encounters, we can easily imagine that everyone would have their own everyday T-environment unlike the

[6] The most optimistic conclusion I have found in the empirical literature on the improvement of deception detection accuracy (when it comes to real-time detection based on behavioural cues) is that '*certain* professions or *certain* subgroups within professions may develop particular sensitivity to *certain kinds* of lies' (O'Sullivan et al. 2009: 536, *emphasis mine*).

T-environment of any other person or even unlike their own T-environment at different times. If this sounds radical, let's take a look at the example Fricker provides.

> ELLA: Ella, a teenager, has a circle of friends in which there is a strong norm of trust and honesty. They very rarely deceive each other — even to the point of preferring honesty to tactfulness. In the situation of Ella and her circle lying is infrequent, and anyone who lies is unpracticed, guilty and embarrassed. The liar shows tell-tale signs and Ella, no fool and perceptually well-attuned, is able to detect them. Then Ella's father gets a new job, and the family moves to a city in another part of the country. Ella goes to a new school with very different social conditions and mores. In her new school, lots of the pupils lie regularly in pursuit of their own selfish purposes and are practiced and proficient dissemblers. Ella retains the perceptual attunedness she previously possessed — her 'narrow' T-method is the same. But in her new environment the old skill is not sufficient for the different and more taxing task of detecting when these cynical streetwise city kids, her new classmates, are lying. (Fricker 2016: 97-8)

Now let's imagine that Ella takes part in an experimental study on deception detection. If it takes place before she moves to the new city, according to Fricker's standards, to actually measure Ella's deception detection skills, the setting of the study should replicate the environment of her old school. But what if the study took place a couple of months after she moved to the new school? With the environment of the new school becoming her new everyday T-environment, her filtering skills gradually attune to the reality of life for the 'cynical streetwise city kids.' To measure her deception detection skills after spending a couple of months in the new environment, the study should replicate the T-environment of the new school. For now, we take into account still relatively broadly characterized T-environments of the old school and the new school but there is no principled reason why we should stop here. Every single student at each of these schools encounters a different set of testifiers (assuming that no one is their own testifier), which might affect the frequency and manner of the false testimony they are exposed to, thus, plausibly, every student has their own T-environment.

Reconstructing such specific individual differences is impossible in an experimental setting. After all, it is the bread and butter of empirical research that an experimental design balances between keeping the environment as natural as possible while simultaneously controlling for variables that might affect the result. What comes close to fulfilling Fricker's expectations are studies devoted to deception detection in intimate relationships: romantic relationships, friendships, or parent-child

relationships (Mccornack and Parks 1986; Evans et al. 2016; Levine and Knapp 2018). Interestingly, many of these studies demonstrate that deception detection accuracy is actually lower in close relationships than between strangers (even though participants are more confident in their judgment). Others demonstrate that it is slightly higher in close relationships, but still no higher than 65% (Levine and Knapp 2018).[7]

Fricker openly discards the assumption that 'the data which show that recipients of testimony are bad at detecting lying concern studies in a very specific experimental setting; but the nature of the findings may nonetheless be such that it is likely that they will carry over to other situations in which testimony is given and received.' (Fricker 2016: 102). Similar assumptions are fairly standard in social psychology. I do not want to imply that there are no problems with the ecological validity of deception detection research, only that there is no reason to flat out dismiss them based on this observation. What is lacking in Fricker's critique, is an argument that real-time deception detection is uniquely environment-dependent and thus virtually impossible to investigate empirically.

I understand and sympathize with Fricker's worry that, while appealing to empirical research in philosophical discussion, we might be tempted to cherry-pick studies which support our points. That is why it is important to, whenever it is available, take into account not only particular studies but also meta-analyses which reveal general tendencies in bodies of research consisting of multiple studies. Meta-analyses available in deception-detection literature point consistently into the direction of only slightly better than chance accuracy of deception detection in general population and fairly limited possibility of its improvement by training (Bond and DePaulo 2006; Hartwig and Bond 2014; Hauch et al. 2016; Sternglanz et al. 2019).

To sum, I think that we should reject Fricker's critique as based on unrealistic demands. Empirical research cited by Michaelian and Shieber (together with further research on deception detection published in the last ten years) makes a very strong case against the real-time effectiveness of filtering. But I also think that we can do even more to demonstrate that filtering is not real-time effective. In the next section, I will take up Fricker's challenge of demonstrating that '[h]umans are constitutionally incapable of learning to respond discriminatingly to testimony.' (2016: 103).

[7] Some of these studies have already been discussed by Shieber (2015: 33-4), yet Fricker does not take them into account.

2.2 Filtering cannot be real-time effective

On Fricker's account (see, e.g. Fricker 1994), the process of forming testimony-based beliefs, has two important features. First, the process takes as an input the receiver's belief about what the speaker said (e.g. *that the speaker said that p*). For the purposes of the present discussion, I will call such beliefs *comprehension-based beliefs*. Second, based on the input it receives, the belief-forming process outputs either a testimonial belief that *p* or no belief at all.

Here is the model of language comprehension underlying Fricker's account: upon hearing or reading an utterance, receivers (i) form comprehension-based beliefs representing the speaker as asserting[8] certain content (*p*), and then (ii) either accept or reject *p* based on the assessment of the speaker's honesty and competence, which leads either to the formation of a corresponding testimony-based belief (that *p)* or no formation of belief. Since the seminal work of a psychologist Daniel Gilbert and his colleagues (Gilbert et al. 1990; Gilbert 1991; Gilbert et al. 1993), this model is often called *the Cartesian model* of language comprehension. If one thinks about filtering in terms of the Cartesian model, it is natural to assume that the filter is 'located at the entrance to our belief box' and that its role is to keep contents of testimony from unreliable (dishonest or incompetent) sources from falling into the box. The Cartesian model is so popular across the current philosophical debate, that one could assume it is the only game in town. But it is not.[9]

In an alternative (*Spinozan)* model, upon comprehending an utterance that *p*, a receiver automatically accepts the content of this utterance and forms a belief that *p*. Later on, they can reject the belief and remove *that p* from their belief box. However, rejection is an additional step, which requires extra time and cognitive resources. Gilbert himself supported the Spinozan model. In a series of experiments, he demonstrated that under cognitive and time pressure, people form beliefs based on comprehended contents that are explicitly identified as false; belief-formation is not optional but automatic and mandatory. If one thinks about filtering in terms of the Spinozan model, one must assume that there is only filtering *ex post*. Nothing could prevent the initial formation of a belief that *p* based on comprehension of an assertion that *p*. This sounds (and indeed is) very radical. What about the comprehension of blatantly false and improbable assertions? Do I automatically form beliefs that Paris is in Germany or that the Earth is flat upon hearing these pieces of information being asserted?

[8] This is the case for assertoric speech. Plausibly, in cases of other speech acts, receivers represent speakers as asking, ordering, etc.
[9] It is worth mentioning that both Michaelian (2010: 403, footnote 9) and Shieber (2015: 38-9 & 93) are aware of this fact.

Thirty years of empirical research on language comprehension since Gilbert's seminal studies suggest that neither the Cartesian nor Spinozan model gives a fully accurate picture of language comprehension (Hasson et al. 2005; Street and Richardson 2015; cf. Kissine and Klein 2013; Grodniewicz 2020, ms). Apparently, *contra* the Cartesian model, a lot of what we comprehend is in fact automatically accepted before the credibility of the speaker is taken into account. Weil et al. (2020), suggest that '… [receivers] consider the credibility of a source only after they have comprehended information and evaluated its consistency with the active memory contents. Accordingly, source credibility might not influence the initial encoding of the information…' (p. 231). However, *contra* the Spinozan model, we do not automatically accept whatever we are being told. In particular, we do not accept contents that are glaringly inconsistent with our active or easily accessible background knowledge. Besides the source-oriented filter commonly discussed in the philosophy of testimony, there is a content-oriented filter which the literature on language comprehension calls *validation*, and which prevents us from believing obviously false information (Richter et al. 2009; Richter 2015; Singer 2019). Validation is sensitive to, e.g. 'violations of factual world knowledge (e.g. *Soft soap is edible*), implausibility (e.g. *Frank has a broken leg. He calls the plumber*), inconsistencies with antecedent text (e.g. *Mary is a vegetarian.... She orders a cheeseburger*), and semantic anomalies (e.g. *Dutch trains are sour*)' (Isberner and Richter 2014: 246).

Therefore, it seems that language comprehension involves two main types of filtering mechanisms: one content-oriented and faster, the other one source-oriented and slower, which are, at least to some extent, independent from one another. Based on these considerations, in (Grodniewicz 2020, ms) I propose a dual-stream model of language comprehension.[10] The first stream is faster and entirely content-oriented. It updates contents of comprehended assertions into our belief box unless they are filtered out by validation. For example, if Liv comprehends Tom's utterance that *p*, this stream processes only the content of the utterance (*p*), and forms a belief that *p*, unless *p* is obviously false to Liv. The second stream is slower and 'source-oriented.' Similarly to the Cartesian model, it operates on representations of contents as produced by a given speaker (e.g. *that Tom asserted that p*). In this stream, some contents are filtered out as originating from a source that the receiver recognizes as unreliable. For example, if Liv knew that Tom is a prolific liar, the content *p* could be filtered out as being asserted by Tom. Crucially, however, it is not the case that at the end of the second stream subjects either form a belief that *p* or no belief at all. The first, purely content-oriented stream is faster. Therefore, if the content of the assertion passes the gatekeeper

[10] Notably, Shieber (2015: 39) also suggests that dual-process models of language comprehension, according to which we acquire many testimony-based beliefs via fast, heuristic routes, are both empirically plausible and problematic for reductionist accounts of testimonial entitlement.

of validation, *that p* is already in the subject's belief box. Monitoring for the trustworthiness and competence of the source can, at most, trigger an attempt of belief revision but does not prevent its formation.

In this model, each stream has its respective filter: validation and source monitoring.[11] Unfortunately, neither of these filters turns out to be real-time effective. Validation filters out only blatantly false and inconsistent information, but it remains virtually helpless against plausible falsehoods (Isberner and Richter 2014; Marsh et al. 2016). Source oriented filtering, on the other hand, not only has all the problems enumerated by Michaelian and Shieber, but can at most trigger an attempt of belief revision.

If this model is correct, we go a long way towards demonstrating that '[h]umans are constitutionally incapable of learning to respond discriminatingly to testimony' (Fricker 2016: 103). While posing this challenge, Fricker focused on the possibility of the improvement of filtering as it is conceptualized in the Cartesian model. But the truth about the mechanisms underlying the comprehension of testimony seems to be more complicated than the Cartesian model suggests. Apparently, the evolved cognitive setup of our comprehension is not optimized for effective discrimination between reliable and unreliable testimony, and there is no reason to think that we can alter this predicament if we just try harder.

To this point, I have argued that filtering is not real-time effective, i.e. the process of filtering, which takes place simultaneously with or right after the process of comprehension, does not allow a particular hearer on a particular occasion to prevent the formation of beliefs based on false testimony. Now I proceed to the second part of my discussion on filtering. It is concerned with its long-term effectiveness.

3. Long-term effectiveness of filtering

The discussion about filtering within the epistemology of testimony is predominantly occupied with the matter of its real-time effectiveness. But even if it is not effective in the real-time, filtering might be long-term effective. As a reminder:

> **Long-term effective filtering (LTEF):** Filtering reliably prevents hearers from forming beliefs based on false testimony because the fact that some hearers sometimes detect

[11] Here, I am not using 'source monitoring' in the sense familiar from the memory literature, i.e. as referring to an ability to attribute particular memory records to their respective sources (cf. Johnson et al. 1993), but as synonymous with 'vigilance towards the source' (cf. Sperber et al. 2010).

> false testimony increases speakers' incentive for honesty and stabilizes the practice of human communication in which deception is risky, costly, and thus not very prevalent.

In this section, I will argue that we have good reasons to assume that LTEF is true. They originate from two different but ultimately converging sources. The first is the evolutionary hypothesis about the stability of human communication championed by Dan Sperber (Sperber et al. 2010; Sperber 2013). The second is the socio-psychological theory — the Truth-Default-Theory — developed by Timothy Levine (2014, 2019b). I will discuss them one by one.

Dan Sperber offers a view very similar to LTEF in his (2013) paper 'Speakers are honest because hearers are vigilant.' This paper is an answer to Kourken Michelian's critique of Sperber's earlier work (Sperber et al. 2010) in which Sperber and his colleagues suggested that, in the evolutionary perspective, filtering[12] is responsible for assuring the stability of human communication. Michaelian (2013), using arguments similar to this which he deployed against Fricker in his (2010) paper, argues that filtering cannot ensure the evolutionary stability of communication, because it is not effective. But Sperber points out that Michaelian's account of filtering is too narrow. Even if not effective in the real-time, filtering is essential in the long run.

This hypothesis arises from two independent, well-established, and fairly minimal observations. Firstly, *some of us* are *occasionally* successful in detecting dishonesty and incompetence,[13] either in the real-time or (substantially more often) with a delay.[14] Secondly, being caught on dishonesty and incompetence triggers social retribution (Dunbar 1996; Dessalles 2007; Sperber and Baumard, 2012). The combination of these two factors puts pressure on speakers:

> Quasi-universal vigilance makes dishonesty less likely to be beneficial in the short run and more likely to be costly in the long run: falsehoods may be disbelieved, and dishonesty may have reputational costs. (Sperber 2013: 69)

Filtering is long-term effective without being real-time effective. It does not enable us to reliably catch false testimony in a particular situation, but it decreases the probability that we will encounter false testimony in the first place.[15]

---

[12] Sperber et al. (2010) use the term *epistemic vigilance* which they define as 'a suite of cognitive mechanisms… targeted at the risk of being misinformed by others' (2010: 359). I will use 'filtering' and 'epistemic vigilance' interchangeably.
[13] Cf. Solbu and Frank's (2019) who argue that we might be collectively effective in catching lies thanks to the fact that 'some individuals are more apt at being good lie detectors' (p. 40).
[14] Cf. Park et al. (2002).
[15] Even though Michaelian (2012, 2013) sharply differentiates his view from Sperber's, I think that, at least to some extent, they can be reconciled. In particular, Michaelian (2012) claims that dishonesty is not very prevalent because

But maybe speakers are naturally inclined to honesty *independently* of hearers' filtering? This seems to be implied by Timothy Levine in his Truth-Default-Theory (TDT) (Levine 2014, 2019b). According to Levine, it is an empirical fact that dishonesty is not very prevalent. As indicated by survey studies conducted in the US, UK, Netherlands, Japan, and Korea (Serota et al. 2010; Halevy et al. 2014; Serota and Levine 2015; see also discussion in Levine 2019b, Chapter 9), '[m]ost communication by most people is honest most of the time' (Levine 2014: 9) while the majority of lies are produced by a few prolific liars. Interestingly, Levine explicitly rejects the speculation about an evolutionary arm-race between speaker's benefiting from deception and hearer's benefiting from deception detection. Thus, at least seemingly, he contradicts Sperber:

> I have heard and read the argument many times that since humans evolved to deceive, we must have evolved the ability to detect deception… I do not think that accepting evolution requires accepting a coevolutionary struggle between the ability to deceive and the ability to *detect deception in real time*. (Levine 2019b: 187*, emphasis mine*)

As we see, Levine rejects the hypothesis that the arms race promotes evolution of the ability to detect deception *in the real-time*. But this is not what Sperber argues for. What he argues for is that the evolutionary arm race forced the receivers to develop a suite of filtering skills, which are effective *in the long-term*: by shaping the social reality of our communicative practice into one in which deception is risky and difficult. This much seems to be at least compatible with Levine's own observations:

> We have created cultures, religions, and socialization that seek to prevent deception… Prevention reduces the prevalence and risk of deception to make the truth-default payoff stronger. It's more efficient to prevent deception than to evolve brains well suited to real-time deception detection. (Levine 2019b: 189)

Here, Levine speaks as if prevention was restricted to social conventions and constraints. But if we look under the hood of his theory, we will see that what he postulates is almost indistinguishable from Sperber's vigilance. The core of TDT is that: it is adaptive for participants in communicatory exchanges 'to operate on a default presumption that what the other person says is basically honest' (Levine 2014: 1). A receiver operating under this presumption remains in what Levine calls a *truth-default state*. However, the state can be abandoned upon encountering a trigger event:

---

deception is costly. He enumerates three main costs of deception: cognitive, psychological, and social. The social cost of deception results from the fact that deception, if detected, is often punished — a possibility that a liar has to always take into account. I suggest that, translating LTEF and Sperber's account into Michaelian's terminology, we could say that filtering is (long-term) effective by increasing the social costs of deception.

> Trigger events include, but are not limited to (a) a projected motive for deception, (b) behavioral displays associated with dishonest demeanor, (c) a lack of coherence in message content, (d) a lack of correspondence between communication content and some knowledge of reality, or (e) information from a third party warning of potential deception. (Levine 2014: 9)

But, of course, trigger events themselves have to be detected somehow to push a receiver out from the default-state.[16] Some kind of 'low-key monitoring' (Sperber 2013: 64) or sensitivity to possible deception has to be active *all the time*, even in the truth-default state.[17]

I conclude that, despite superficial differences, Levine and Sperber's theories are compatible and simultaneously supported by the evolutionary hypothesis articulated by Sperber, and the empirical data about the prevalence of honesty collected by Levine. Moreover, both of them support LTEF.

Admittedly, the empirical grounds on which I base my defence of LTEF are less robust than the ones I have discussed in my rejection of RTEF in the previous section.[18] While evaluating RTEF, we could appeal to a plethora of empirical studies directly investigating real-time performance of testimony-receivers. The case of LTEF is much more complicated.

Firstly, it is difficult to gather reliable data regarding the sheer volume of false testimony in our everyday lives. The closest we get to it, is by appealing to studies — used by Levine to support TDT — based on reports of lying frequency (e.g. Serota et al. 2010; Halevy et al. 2014; Serota and Levine 2015). These studies suggest that people tell on average 1-2 lies per day, with big lies being significantly less prevalent (less than 0.5 per day) than little white lies, such as exaggerated compliments.[19] Nevertheless, even under the assumption that, on average, we lie 1-2 times per day, it is virtually impossible to estimate the ratio of dishonest to honest testimony in our social environment because it would require estimating how many times per day, on average, we say something true.[20] Finally, and most importantly, it is difficult to demonstrate that the relatively small amount of lies in our environment is *actually* a result of the long-term effectiveness of filtering. Luckily, a

[16] A similar picture is suggested by Lipton (2007) in his default-trigger model of testimony.
[17] Obviously, once receivers abandon the truth-default state upon detecting a trigger event, they *do not* become able to effectively detect deception in the real-time. Instead, they become suspicious and motivated to look for further evidence of the speaker's dishonesty. This *might* lead — typically with a delay — to identifying given content as false: 'most lies are detected after-the-fact based on either confessions or the discovery of some evidence showing that what was said was false.' (Levine 2014: 6).
[18] I am grateful to an anonymous reviewer for pressing me to clarify this.
[19] Approximately 80% of all lies in the U.K. population were little white lies (Serota and Levine 2015: 12).
[20] We can assume that, e.g. a person whose work requires frequent communication with their co-workers can easily communicate hundreds of true statements per day.

huge meta-study conducted recently by Johannes Abeler and his colleagues (2019) makes this hypothesis more probable than ever before.

Abeler and his collaborators combined data from 429 experiments across 90 papers, which jointly involved more than 44000 participants across 47 countries. The setup of all these studies was almost identical: subjects conduct a random draw (e.g. a roll of six-sided die) and report the outcome to the experimenter who pays them a monetary reward equal to the number which they reported. The meta-study showed that even in such a restricted and gamified context 'subjects forego about three-quarters of the potential gains from lying' (Abeler et al. 2019: 1123). Crucially for my argument, according to the authors, the only models which could account for such results are ones that assume subjects' desire to be honest *jointly* with their desire to appear honest:

> … our results suggest that a preference for being seen as honest and a preference for being honest are the main motivations for truth-telling. Finally, policy interventions that rely on voluntary truth-telling by some participants could be very successful, in particular if it is made hard to lie while keeping a good reputation. (Abeler et al. 2019: 1123)

Therefore, despite all the problems enumerated above, LTEF is not only hypothetically plausible but also relatively well supported by empirical data.[21]

Let us sum up what we have established to this point. Even though filtering is not real-time effective (RTEF is false), it is effective in the long-term (LTEF is true) and thus filtering *is* effective

[21] An anonymous reviewer points out that Shieber (2015: 82-4) offers empirical evidence that undermines the relationship between the risk of reputation loss and truthfulness, and thus threatens LTEF. I will briefly discuss three main pieces of evidence provided by Shieber. Firstly, he cites Olszewski and Sandroni's (2011) paper in which, based on a game-theoretic model, they demonstrate that falsification (as construed by Popper) is not an effective strategy for distinguishing between reliable and unreliable experts. However, in the same paper, Olszewski and Sandroni demonstrate that a different strategy, i.e. refutation, *is* effective in the same context (2011: 792). Given that, I believe that their paper does not provide sufficient evidence that it is impossible to distinguish between reliable and unreliable experts, but only that it is impossible to do it by appealing to the criterion of Popperian falsification. Secondly, Shieber discusses Tetlock's (2005) research on expert political judgment, which suggests that most vocal and popular political experts are often unreliable when it comes to factual information and, at the same time, their predictions are rarely checked for accuracy. It should be noted, however, that Tetlock himself seems to sympathize with LTEF-style solutions when he says: 'But I do still believe it possible to raise the quality of debate by tracking the quality of claims and counterclaims that people routinely make…' (Tetlock 2005: 218). Finally, Shieber argues that the state of contemporary science (especially the replication crisis) can be seen as evidence that 'reliance on reputation and sanctions is not sufficient to ensure truthfulness' (Shieber 2015: 83). This is an interesting topic that calls for careful discussion considering all the different kinds of pressures (economic, social, etc.) which shape the environment of modern scientific research. However, it is unclear whether these observations undermine LTEF. Despite its shortcomings, such as the difficulty of publishing replication studies in top scientific journals, or the citation bias, one can still argue that the environment of modern scientific research seems to remain one in which deception is risky and costly, and thus not very prevalent (e.g. Steen 2011; Lüsher 2013). After all, my argument for LTEF *does not* require long-term filtering to ensure perfect or near-perfect truthfulness.

(EF is true). It is not real-time effective because it does not reliably block false contents of comprehended testimonies from falling into our belief box. There are two reasons why this is not the case. First, we are really bad at identifying unreliable testimonies on the fly. Second, the very idea that each time we comprehend a given assertoric utterance, we are free to either accept or reject it, is based on an idealized and inaccurate picture of linguistic comprehension. Instead, we typically accept what is said in comprehended assertoric utterances upon only minimal, content-oriented filtering (validation). Nevertheless, filtering is not pointless. Quite the opposite, in the long-term filtering shapes and sustains the practice of human communication in which dishonesty is both costly and risky, and thus far less prevalent than it would have been was filtering absent. In the next section, I spell out the consequences of this picture for the debate about the epistemology of testimony.

4. Strong social anti-reductionism about testimonial entitlement

Are we by default entitled to believe what we are being told? To answer 'no' is to profess reductionism about testimonial entitlement; to answer 'yes' is to profess anti-reductionism. In this final section, I suggest that the picture of effective filtering outlined above supports the following version of anti-reductionism about testimonial entitlement:

> **Strong social anti-reductionism (SSAR):** Hearers are prima facie entitled to form beliefs based on what speakers assert because filtering is long-term effective.

In calling this view 'strong' I follow Mona Simion and Christoph Kelp's (2018; Simion 2020) distinction into *moderate* and *strong* anti-reductionism. According to strong anti-reductionism, we are prima facie (absent defeaters) entitled to believe *whatever* we are being told. According to moderate anti-reductionism, on the other hand, some additional condition has to be met for our testimony-based beliefs to enjoy prima facie entitlement. SSAR is a version of strong anti-reductionism because it does not impose any special condition a hearer has to meet to be prima facie entitled to their testimony-based beliefs. Each and every one of us is prima facie entitled to believe what other people say not because of our individual merits and skills, but due to the very fact of participating in human communication — a social practice shaped and sustained by long-term effectiveness of filtering.

SSAR is a close kin of Peter Graham's (2010) theory, according to which hearers are prima facie pro tanto entitled to testimonial beliefs formed by the process of comprehension-with-filtering. Graham argues that:

> …comprehension-with-filtering has forming true beliefs reliably as a function, not comprehension neat, comprehension taken alone. It's the filtering, or so I argue, that is for producing a sufficiently high truth ratio. (Graham 2010: 170)

Graham does not appeal to the distinction into real-time and long-term effectiveness of filtering, but he seems to be aware that filtering is long-term effective: 'Filtering not only filters out false or misleading assertions, it also provides an incentive for speakers to not mislead in the first place.' (Graham 2010: 173).[22] What I think is misleading in Graham's account, is the very opposition between the process of comprehension taken alone and the process of comprehension-with-filtering. In human communication, there is no comprehension *without* filtering. To see this, we have to look at filtering simultaneously from the individual and social perspective.

Starting with the individual perspective. Filtering takes many forms and it is plausible that some basic filtering mechanisms (such as validation discussed in Section 2.2) accompany all instances of comprehension. Two possible objections come to mind. First, maybe validation should not count as a form of filtering because it prevents only the acquisition of beliefs in glaring disagreement with our background knowledge, so it very rarely prevents us from forming false testimony-based beliefs. Obviously, this objection does not go through. As we demonstrated in our discussion about real-time effectiveness of filtering, even the most sophisticated filtering mechanisms are not effective in real-time, i.e. do not reliably prevent us from forming false beliefs based on a particular instance of testimony in a particular situation.

But maybe — and this is a possible second objection — very young children's comprehension is an example of comprehension without filtering? There are two things to be said in response to this objection. Firstly, given that filtering consists of a multitude of mechanisms: some very sophisticated (like paying attention to someone's social demeanour), some much more primitive (like detecting incongruences between what one says and what we see), it is quite likely that some basic forms of filtering appear in children's development simultaneously with the ability to comprehend language. Again, to qualify as a filtering mechanism, a mechanism does not have to be real-time effective. However, even if it would turn out that, at an early stage of development, children are able to comprehend language, but without any sort of filtering, this does not yet mean that their comprehension is comprehension *without* filtering. It is comprehension without filtering only from

[22] This is why Simion and Kelp's (2018; Simion 2020) critique of Graham's view misses the mark. They argue that the filtering requirement is redundant because filtering is not effective, but they support this claim by appeal only to Michaelian's arguments against the *real-time* effectiveness of filtering (Simion and Kelp 2018: 7).

the individual perspective, i.e. they *themselves* cannot perform filtering. But we cannot forget about the social perspective.

Filtering is something we do for each other at least as much as we do it for ourselves. We shape our common social environment by monitoring it. Even if small children themselves do not perform filtering, they operate in an environment shaped by the filtering performed by adults. Once we realize this, we see that in our everyday linguistic practices, there is really no comprehension without filtering.

This picture bears important similarities to the way of treating the problem of childhood testimony by Sanford Goldberg (2007).[23] Goldberg suggests that, while assessing whether a child acquired knowledge by comprehending a given testimony, we should take into consideration the context in which it happened, especially the role played by the child's adult guardians. According to Goldberg, adult caretakers actively monitor their child's environment and thus 'enhance the reliability of a good many of the beliefs that are elicited by the child's encounters with testimony' (2007: 221). While Goldberg focuses on children as the beneficiaries of the monitoring done by others, I think that it is not only children but all members of the linguistic community. Taken separately, and from the perspective of a particular instance of reception of testimony, we are all nearly child-like vulnerable to acquiring false beliefs. Our ability to detect deception in real-time is only slightly better than chance. However, being vigilant by default, every adult language user constantly takes care of the whole linguistic community. This is why all our beliefs based on the comprehension of testimony enjoy prima facie entitlement. We are prima facie entitled to believe what we are being told not because we are likely to recognize if a particular utterance is false, but because in the communicative environment shaped by routine monitoring, dishonesty is risky, costly, and thus not very prevalent.

Finally, SSAR is compatible with Simion's (2020) version of strong anti-reductionism, i.e. *Testimonial contractarianism,* but it points at a more fundamental source of our testimonial entitlement. According to *Testimonial contractarianism*, we are prima facie entitled to believe whatever we are being told because, in virtue of the social contract in play, speakers are by default compliant with the norms governing speech acts (in particular, the *knowledge norm* of assertion: one should only assert that *p* if one knows that *p*). Speakers are by default compliant with these norms because this is the rational thing to do if one is not oriented towards one's straightforward and immediate self-interest, but towards long-term, constrained self-interest (Simion 2020: 23). If we were oriented towards

[23] But notice that Goldberg focuses on the acquisition of knowledge, while I focus only on prima facie entitlement.

maximizing only straightforward self-interests, it would be rational to lie whenever a given lie might bring about immediate benefits. But since we are oriented towards constrained self-interest, the fact that our lies might be detected (immediately or with a delay; by the receiver themselves or by someone else in the community) changes what is the rational thing to do. Social reputation, which we risk if we are caught on unreliable testimony, is often more valuable than whatever we gain by lying in a particular situation. In short, what makes it beneficial for speakers to comply with the norms governing speech acts and thus, with a social contract like the one proposed by Simion, is that filtering is long-term effective. The normativity of language use described by Simion is grounded in the psychology, sociology, and epistemology of filtering described above.

Before I conclude, let me briefly address two additional objections that can be raised against my account.[24] Firstly, one can wonder whether there is a particular threshold of reliability of long-term filtering necessary to support SSAR. Would 70% be enough? What about 75%? I think that it is neither possible nor necessary to establish such a precise threshold. This would require, among other things, assessing what percentage of all testimony we encounter is false testimony (a problem I have mentioned in the previous section). Again, given the sheer amount of information we exchange every day, I find it very plausible that on average the percentage is significantly smaller than 30% or even 25%. This is sufficient for SSAR. Ultimately, all that strong anti-reductionism about testimonial entitlement tries to establish is that 'the default position for hearers is entitlement to believe, just like the default position for pedestrians is to cross the street on a green light' (Simion 2020: 20).

Secondly, my discussion throughout this paper focused mostly on filtering of dishonest testimony, which raises the question whether it is also applicable to incompetent testimony — equally relevant to SSAR. I do think that my solution is equally applicable to the problem of incompetence. Just as dishonesty is often beneficial in the short term, but has significant long-term costs, so does incompetence. An incompetent testifier does not have to waste energy on collecting reliable information and fact checking. They can offer apparent epistemic goods fast and cheap. Moreover, most receivers will not be able to recognize the incompetent testimony to be false in the real time, especially, if the testifier is sufficiently confident (cf. Shieber 2015: 42-3). However, in the long run, incompetent testifiers risk losing their social status just like dishonest ones. Sooner or later, the original receiver of testimony or some other member of the community, may be in position to

[24] I am grateful to anonymous reviewers for pushing me to address these issues.

detect the falsehood and hypothesize that the source was either dishonest or incompetent — in either case, a suspicious source of future testimony.

5. Conclusions

So, is filtering effective? In this paper, I have argued that we should look at this question from a broader perspective. Filtering is not real-time effective because it does not allow us to respond discriminatingly to particular instances of testimony. We are really bad at online deception detection. But filtering is long-term effective. It prevents us from forming a large number of false testimonial beliefs not by turning each of us into a high-functioning polygraph, but by turning the social environment of human communication into one in which such polygraphs are not required.

This way of looking at the effectiveness of filtering allows us to reconsider the role that language comprehension plays in the acquisition of testimony-based beliefs. Firstly, it allows us to come to terms with a growing body of empirical research, suggesting that we are not free to either accept or reject whatever we comprehend. It is quite likely that acceptance is the default reaction to comprehended content, and thus real-time effective filtering is simply impossible. Secondly, on the ground of the debate about testimonial entitlement, these considerations support a version of strong anti-reductionism, i.e. the view according to which we are prima facie entitled to believe whatever we are being told.[25]

Funding

The research has been supported by a grant from the Priority Research Area 'Society of the Future' under the Strategic Programme 'Excellence Initiative' at Jagiellonian University.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for Truth-Telling. Econometrica, *87*(4), 1115–1153. https://doi.org/10.3982/ECTA14673

Adler, J. E. (1994). Testimony, Trust, Knowing: Journal of Philosophy, 91(5), 264–275. https://doi.org/10.2307/2940754

[25] I thank Bartłomiej Czajka, Anna Drożdżowicz, Manuel García-Carpintero, Grzegorz Gaszczyk, Sanford Goldberg, Peter J. Graham, Josep Macià, Neri Marsili, Michele Palmira, Andrew Peet, Josefa Toribio, and two anonymous reviewers for their comments and advice.

Audi, R. (1997). The Place of Testimony in the Fabric of Knowledge and Justification. American Philosophical Quarterly, 34(4), 405–422.

Audi, R. (2006). Testimony, Credulity, and Veracity. In J. Lackey & E. Sosa (Eds.), The Epistemology of Testimony (pp. 25–46). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276011.003.0002

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of Deception Judgments. Personality and Social Psychology Review, 10(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Bond, C. F., Kahler, K. N., & Paolicelli, L. M. (1985). The miscommunication of deception: An adaptive perspective. Journal of Experimental Social Psychology, 21(4), 331–345. https://doi.org/10.1016/0022-1031(85)90034-4

Burge, T. (1993). Content preservation. Philosophical Review, 102(4), 457–488.

Burge, T. (1997). Interlocution, Perception, and Memory. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 86(1), 21–47.

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

DePaulo, B. M., & Pfeifer, R. L. (1986). On-the-Job Experience and Skill at Detecting Deception1. Journal of Applied Social Psychology, 16(3), 249–267. https://doi.org/10.1111/j.1559-1816.1986.tb01138.x

Dessalles, J.-L. (2007). Why we talk: The evolutionary origins of language. Oxford University Press.

Dunbar, R. I. M. (1996). Grooming, gossip, and the evolution of language. Harvard University Press.

Evans, A. D., Bender, J., & Lee, K. (2016). Can parents detect 8- to 16-year-olds' lies? Parental biases, confidence, and accuracy. Journal of Experimental Child Psychology, 147, 152–158. https://doi.org/10.1016/j.jecp.2016.02.011

Feeley, T. H., & Young, M. J. (2000). Self-reported cues about deceptive and truthful communication: The effects of cognitive capacity and communicator veracity. Communication Quarterly, 48(2), 101–119. https://doi.org/10.1080/01463370009385585

Fricker, E. (1987). The epistemology of testimony. Proceedings of the Aristotelian Society, Supplementary Volumes, 61, 57–83.

Fricker, E. (1994). Against Gullibility. In A. Chakrabarti & B. K. Matilal (Eds.), Knowing from Words. Kluwer Academic Publishers.

Fricker, E. (1995). Telling and Trusting: Reductionism and Anti-Reductionism in the Epistemology of Testimony. Mind, 104(414), 393–411. https://doi.org/10.1093/mind/104.414.393

Fricker, E. (2006). Varieties of Anti-Reductionism About Testimony? A Reply to Goldberg and Henderson. Philosophy and Phenomenological Research, 72(3), 618–628. https://doi.org/10.1111/j.1933-1592.2006.tb00587.x

Fricker, E. (2016). Unreliable testimony. In B. P. McLaughlin & H. Kornblith (Eds.), Goldman and His Critics (pp. 88–123). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118609378.ch5

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. Journal of Personality and Social Psychology, 65(2), 221–233. https://doi.org/10.1037//0022-3514.65.2.221

Gilbert, D. T. (1991). How mental systems believe. American Psychologist, 46(2), 107–119. https://doi.org/10.1037/0003-066X.46.2.107

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. Journal of Personality and Social Psychology, 59(4), 601–613. https://doi.org/10.1037/0022-3514.59.4.601

Goldberg, S. C. (2006). Reductionism and the Distinctiveness of Testimonial Knowledge. In J. Lackey & E. Sosa (Eds.), The Epistemology of Testimony (pp. 127–141). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276011.003.0007

Goldberg, S. C. (2007). Anti-Individualism: Mind and Language, Knowledge and Justification. http://dx.doi.org/10.1017/CBO9780511487521

Goldberg, S. C. (2010). Relying on others: An essay in epistemology. Oxford University Press.

Goldberg, S. C. (2014). Interpersonal Epistemic Entitlements. Philosophical Issues, 24(1), 159–183. https://doi.org/10.1111/phis.12029

Goldman, A. I. (1999). Knowledge in a social world.

Graham, P. J. (2006). Liberal Fundamentalism and Its Rivals. In J. Lackey & E. Sosa (Eds.), The Epistemology of Testimony (pp. 93–114). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276011.003.0005

Graham, P. J. (2010). Testimonial Entitlement and the Function of Comprehension. In A. Haddock, A. Millar, & D. Pritchard (Eds.), Social Epistemology (pp. 148–174). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199577477.003.0008

Graham, P. J. (2015). Epistemic Normativity and Social Norms. In D. K. Henderson & J. Greco (Eds.), Epistemic Evaluation (pp. 246–273). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199642632.003.0011

Greco, J. (2015). Testimonial Knowledge and the Flow of Information. In D. K. Henderson & J. Greco (Eds.), Epistemic Evaluation (pp. 274–290). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199642632.003.0012

Grodniewicz, J. P. (ms). The Representational Structure of Linguistic Understanding

Grodniewicz, J. P. (2020). Themes in linguistic understanding. Cognition and epistemology [Ph.D. Thesis, Universitat de Barcelona]. In *TDX (Tesis Doctorals en Xarxa)*. http://www.tdx.cat/handle/10803/670332

Halevy, R., Shalvi, S., & Verschuere, B. (2014). Being Honest About Dishonesty: Correlating Self-Reports and Actual Lying. Human Communication Research, 40(1), 54–72. https://doi.org/10.1111/hcre.12019

Hartwig, M., & Bond, C. F. (2014). Lie Detection from Multiple Cues: A Meta-analysis. Applied Cognitive Psychology, 28(5), 661–676. https://doi.org/10.1002/acp.3052

Hasson, U., Simmons, J. P., & Todorov, A. (2005). Believe It or Not: On the Possibility of Suspending Belief. Psychological Science, 16(7), 566–571. https://doi.org/10.1111/j.0956-7976.2005.01576.x

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2016). Does Training Improve the Detection of Deception? A Meta-Analysis. Communication Research, 43(3), 283–343. https://doi.org/10.1177/0093650214534974

Hume, D. (1975). David Hume: Enquiries Concerning Human Understanding and Concerning the Principles of Morals (Third Edition) (P. H. Nidditch, Ed.). Oxford University Press. https://doi.org/10.1093/actrade/9780198245353.book.1

Isberner, M.-B., & Richter, T. (Eds.). (2014). Comprehension and Validation: Separable Stages of Information Processing? A Case for Epistemic Monitoring in Language Comprehension. In Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences (pp. 245–276). The MIT Press.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. Psychological Bulletin, *114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Kissine, M., & Klein, O. (2013). Models of communication, epistemic trust and espistemic vigilance. In J. P. Forgas, O. Vincze, & J. László (Eds.), Social cognition and communication (pp. 139–154). New York: Psychology Press. 10.4324/9780203744628

Köhnken, G. (1987). Training police officers to detect deceptive eyewitness statements: Does it work? Social Behaviour.

Kraut, R. E., & Poe, D. B. (1980). Behavioral roots of person perception: The deception judgments of customs inspectors and laymen. Journal of Personality and Social Psychology, 39(5), 784–798. https://doi.org/10.1037/0022-3514.39.5.784

Lackey, J. (2006). It Takes Two to Tango: Beyond Reductionism and Non-Reductionism in the Epistemology of Testimony. In J. Lackey & E. Sosa (Eds.), The Epistemology of Testimony (pp. 160–182). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276011.003.0009

Levine, T. R. (2010). A few transparent liars. Communication Yearbook, 34, 40–61.

Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. Journal of Language and Social Psychology, 33(4), 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R. (2019a). An Overview of Detecting Deceptive Communication. In T. Docan-Morgan (Ed.), The Palgrave Handbook of Deceptive Communication (pp. 289–301). Springer International Publishing. https://doi.org/10.1007/978-3-319-96334-1_15

Levine, T. R. (2019b). Duped: Truth-default theory and the social science of lying and deception. The University of Alabama Press.

Levine, T. R., & Knapp, M. L. (2018). Lying and deception in close relationships. In The Cambridge handbook of personal relationships, 2nd ed (pp. 329–340). Cambridge University Press. https://doi.org/10.1017/9781316417867.026

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the 'veracity effect.' Communication Monographs, 66(2), 125–144. https://doi.org/10.1080/03637759909376468

Lipton, P. (1998). The epistemology of testimony. Studies in History and Philosophy of Science Part A, 29(1), 1–31. https://doi.org/10.1016/S0039-3681(97)00022-8

Lipton, P. (2007). Alien Abduction: Inference to the Best Explanation and the Management of Testimony. Episteme, 4(3), 238–251. https://doi.org/10.3366/E1742360007000068

Lüscher, T. F. (2013). The codex of science: Honesty, precision, and truth--and its violations. European Heart Journal, *34*(14), 1018–1023. https://doi.org/10.1093/eurheartj/eht063

Lyons, J. (1997). Testimony, induction and folk psychology. Australasian Journal of Philosophy, 75(2), 163–178. https://doi.org/10.1080/00048409712347771

Marsh, E. J., Cantor, A. D., & M. Brashier, N. (2016). Believing that Humans Swallow Spiders in Their Sleep: False Beliefs as Side Effects of the Processes that Support Accurate Knowledge. In B. H. Ross (Ed.), Psychology of Learning and Motivation (Vol. 64, pp. 93–132). Academic Press. https://doi.org/10.1016/bs.plm.2015.09.003

Mccornack, S. A., & Parks, M. R. (1986). Deception Detection and Relationship Development: The Other Side of Trust. Annals of the International Communication Association, 9(1), 377–389. https://doi.org/10.1080/23808985.1986.11678616

Michaelian, K. (2010). In defence of gullibility: The epistemology of testimony and the psychology of deception detection. Synthese, 176(3), 399–427. https://doi.org/10.1007/s11229-009-9573-1

Michaelian, K. (2012). (Social) Metacognition and (Self-)Trust. Review of Philosophy and Psychology, 3(4), 481–514. https://doi.org/10.1007/s13164-012-0099-y

Michaelian, K. (2013). The evolution of testimony: Receiver vigilance, speaker honesty and the reliability of communication. Episteme, 10(1), 37–59. https://doi.org/10.1017/epi.2013.2

Millikan, R. G. (1984). Language, Thought and Other Biological Categories. MIT Press.

Olszewski, W., & Sandroni, A. (2011). Falsifiability. American Economic Review, *101*(2), 788–818. https://doi.org/10.1257/aer.101.2.788

O'Sullivan, M., Frank, M. G., Hurley, C. M., & Tiwana, J. (2009). Police lie detection accuracy: The effect of lie scenario. Law and Human Behavior, 33(6), 530–538. https://doi.org/10.1007/s10979-008-9166-4

Park, H. S., Levine, T., McCornack, S., Morrison, K., and Ferrara, M. (2002). How people really detect lies. Communication Monographs, 69(2):144–157. https://doi-org.sire.ub.edu/10.1080/714041710

Reid, T. (1970). An inquiry into the human mind (T. Duggan, Ed.). University of Chicago Press.

Richter, T. (2015). Validation and Comprehension of Text Information: Two Sides of the Same Coin. Discourse Processes, 52(5–6), 337–355. https://doi.org/10.1080/0163853X.2015.1025665

Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. Journal of Personality and Social Psychology, 96(3), 538–558. https://doi.org/10.1037/a0014038

Serota, K. B., & Levine, T. R. (2015). A Few Prolific Liars: Variation in the Prevalence of Lying. Journal of Language and Social Psychology, 34(2), 138–157. https://doi.org/10.1177/0261927X14528804

Serota, K., Levine, T., & Boster, F. (2010). The Prevalence of Lying in America: Three Studies of Self-Reported Lies. Human Communication Research, 36, 2–25. https://doi.org/10.1111/j.1468-2958.2009.01366.x

Shieber, J. (2012). Against Credibility. Australasian Journal of Philosophy, 90(1), 1–18. https://doi.org/10.1080/00048402.2011.560953

Shieber, J. (2015). Testimony: A philosophical introduction. Routledge, Taylor & Francis Group.

Simion, M. (2020). Testimonial contractarianism: A knowledge-first social epistemology. Noûs. https://doi.org/10.1111/nous.12337

Simion, M., & Kelp, C. (2018). How to be an anti-reductionist. Synthese, 197(7), 2849–2866. https://doi.org/10.1007/s11229-018-1722-y

Singer, M. (2019). Challenges in Processes of Validation and Comprehension. Discourse Processes, 56(5–6), 465–483. https://doi.org/10.1080/0163853X.2019.1598167

Solbu, A., & Frank, M. G. (2019). Lie Catchers: Evolution and Development of Deception in Modern Times. In The Palgrave Handbook of Deceptive Communication (pp. 41–66). Springer.

Sperber, D. (2013). Speakers are honest because hearers are vigilant: Reply to Kourken Michaelian. Episteme, 10(01), 61–71. https://doi.org/10.1017/epi.2013.7

Sperber, D., & Baumard, N. (2012). Moral Reputation: An Evolutionary and Cognitive Perspective. Mind & Language, 27(5), 495–518. https://doi.org/10.1111/mila.12000

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

Steen, R. G. (2011). Retractions in the scientific literature: Is the incidence of research fraud increasing? Journal of Medical Ethics, *37*(4), 249–253. https://doi.org/10.1136/jme.2010.040923

Sternglanz, R. W., Morris, W. L., Morrow, M., & Braverman, J. (2019). A Review of Meta-Analyses About Deception Detection. In T. Docan-Morgan (Ed.), The Palgrave Handbook of Deceptive Communication (pp. 303–326). Springer International Publishing. https://doi.org/10.1007/978-3-319-96334-1_16

Street, C. N. H., & Richardson, D. C. (2015). Descartes Versus Spinoza: Truth, Uncertainty, and Bias. Social Cognition, 33(3), 227–239. https://doi.org/10.1521/soco.2015.33.2.2

Tetlock, P. E. (2005). Expert political judgment: How good is it? How can we know? Princeton Univ. Press.

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. Legal and Criminological Psychology, 9(2), 159–181. https://doi.org/10.1348/1355325041719356

Weil, R., Schul, Y., & Mayo, R. (2020). Correction of evident falsehood requires explicit negation. Journal of Experimental Psychology: General, *149*(2), 290–310. https://doi.org/10.1037/xge0000635

Copernicus Center for Interdisciplinary Studies, Jagiellonian University, Poland