The Ethics and Epistemology of Trust

Trust is a topic of longstanding philosophical interest. It is indispensable to every kind of coordinated human activity, from sport to scientific research. Even more, trust is necessary for the successful dissemination of knowledge, and by extension, for nearly any form of practical deliberation and planning. Without trust, we could achieve few of our goals and would know very little. Despite trust's fundamental importance in human life, there is substantial philosophical disagreement about what trust is, and further, how trusting is normatively constrained and best theorized about in relation to other things we value. This entry is divided into three sections, which explore key (and sometimes interconnected) ethical and epistemological themes in the philosophy of trust: (1) The Nature of Trust; (2) The Normativity of Trust, and (3) The Value of Trust.

Table of Contents

1. The Nature of Trust

What *is* trust? To a very first approximation, trust is an attitude or a hybrid of attitudes (e.g., optimism, hope, belief, etc.) toward a trustee, that involves some (non-negligible) vulnerability to betrayal on behalf of the truster. This general remark, of course, does not take us very far. For example, we may ask: *what kind* of attitude (or hybrid of attitudes) is trust, exactly? Suppose, for example, that (as some philosophers of trust maintain) trust requires an attitude of *optimism*. Even if that is right, getting a grip on trust requires a further conception of what the truster, *qua* truster, must be optimistic *about*. One standard answer here proceeds as follows: trust (at least, in the paradigmatic case of interpersonal trust) involves some form of optimism that the trustee will *take care of things as we have entrusted them*. In the special case of trusting the testimony of another—a topic at the center of the epistemology of trust—this will involve at least some form of optimism that the speaker is living up to her expectations as a testifier, e.g., that speaker knows what she says (e.g., Simion and Kelp Forthcoming, 2020a) or, more weakly, is telling the truth.

Even at this level of specificity, though, the nature of trust remains fairly elusive. Does trusting involve (for example) *merely* optimism that that the trustee will take care of things as entrusted, or does it also involve, for instance, optimism that the trustee will do so compresently with certain beliefs, non-doxastic attitudes, emotions or motivations on the part of the trustee, such as with *goodwill* (e.g., Baier 1986; Jones 1996). Moreover, and apart from such positive characterizations of trust, does trust have also a *negative condition* to the effect that one fails to genuinely trust another if one—past some threshold of vigilance—*monitors* the trustee (or otherwise, reflects critically on the trust relationship so as to attempt to minimize risk)?

These are among the questions that occupy philosophers working on the nature of trust. In this section, we will explore four subthemes aimed at clarifying trust's nature: these concern (a) the distinction between trust and reliance; (b) two-place vs three-place trust; (c) doxastic versus non-doxastic conditions on trust; (d) deception detection and monitoring.

a. Reliance vs. Interpersonal Trust

Reliance is ubiquitous. You rely on the weather not to suddenly drop by 20 degrees, leaving you shivering; you rely on your chair to not give out, causing

you to tumble to the floor. In these cases, are you trusting the weather and trusting your chair, respectively? Plausibly not. This is so even though, in each case, you are depending on these things in a way that leaves you potentially vulnerable.

The idea that trust is a kind of dependence that does not reduce to *mere* reliance (of the sort that might be apposite to things like chairs and the weather) is a widely accepted. According to Annette Baier (1986: 244) the crux of the difference is that trust involves relying on another not just to take care of things any old way (e.g., out of fear, begrudgingly, accidentally, etc.) but rather that they do so out of *goodwill* toward the truster; relatedly, a salient kind of vulnerability one subjects oneself to in trusting is vulnerability to the limits of that goodwill. On this way of thinking, then, you are not *trusting* someone if you (for instance) rely on that person to act in a characteristically self-centred way, even if you depend on them to do so, and even if you fully expect them to do so.

Katherine Hawley (2019) rejects the idea that what distinguishes trust from mere reliance has anything to do with the trusee's motives or goodwill. Instead, on her account, the crucial difference is that in cases of trust, but not of mere reliance, a commitment on the part of the trustee must be in place. Consider a situation in which you reliably bring too much lunch to work, because you are a bad judge of quantities, and I get to eat your leftovers. My attitude to you in this situation is one of reliance, but not trust; in Hawley's view, that is because you have made no commitment to provide me with lunch. But if we adapt the case so as to suggest commitment, it starts to look more like a matter of trust. Suppose we enjoy eating together regularly, you describe your plans for the next day, I say how much I'm looking forward to it, and so on. To the extent that this involves a commitment on your part, it seems reasonable for me to feel betrayed and expect apologies if one day you fail to bring lunch and I go hungry.

If it is right that trust differs in important ways from mere reliance, then a consequence is that while reliance is something we can have toward people (when we merely depend on them) as well as toward objects (e.g., when we depend on gravity and the weather), not just *anything* can be genuinely trusted. Karen Jones (1996) captures this point, one that circumscribes people as the fitting objects of genuine trust, as follows:

> One can only trust things that have wills, since only things with wills can have goodwills—although having a will is to be given a generous interpretation so as to include, for example, firms and government bodies. Machinery can be relied on, but only agents, natural or artificial, can be trusted (1996: 14)

If, as the foregoing suggests, the trust relationships is best understood as a special subset of reliance relationships, should we also expect the appropriate attitudes toward *misplaced* trust to be a subset of a more general attitude-type we might have in response to misplaced *reliance*?

Katherine Hawley (2014) thinks so. As she puts it, misplaced *trust* warrants a feeling of *betrayal*. But the same is not so for misplaced (mere) reliance. Suppose, to draw from an examples she offers (e.g., 2014: 2) that a shelf you rely on to support a vase gives out; it would be inappropriate, Hawley maintains, to feel *betrayed*, even if a more general attitude of (mere) *disappointment* befits such misplaced reliance. Misplaced trust, by contrast, befits a feeling of betrayal.

In contrast with the above thinking, according to which disanalogies between trust and mere reliance are taken to support *distinguishing* trust from reliance, some philosophers have taken a more permissive approach to trust, by distinguishing between *two senses* of trust that differ with respect to their similarities with mere reliance.

Paul Faulkner (e.g., 2011: 246; cf., McMyler 2011) for example, distinguishes between what he calls *predictive* and *affective* trust. Predictive trust involves merely reliance in conjunction with a belief that the trustee will take care of things (viz., a prediction). Misplaced predictions warrant disappointment, not betrayal, and so predictive trust (like mere reliance) cannot be betrayed. Affective trust, according to Faulkner, involves—by contrast-—along with reliance a kind of *normative expectation* to the effect that the trustee (i) ought to prove dependable; and that they (ii) will prove dependable for that reason. On this view, it is affective trust that is uniquely subject to betrayal, even though predictive trust which is not is a genuine variety of trust.

b. Two-place and Three-place Trust

The distinction between two- and three-place-trust, first drawn by Horsburgh (1960) is meant to capture a simple idea: sometimes when we trust someone, we trust them to do some particular thing (e.g., Horsburgh 1960; Holton 1994; Hardin 1992) e.g., you might trust your neighbour to water your plant while you're away on holiday but *not* to look after your daughter. This is three-place trust, e.g., with an infinitival component (schematically: A trusts B to X). Not all trusting fits this schema. You might also simply *trust your neighbour* generally (schematically: A trusts B) and in a way that does not involve any particular task in mind. Three- and two-place trust are thus different in the sense that the object of trust is specified in the former case but not in the latter.

While there is nothing philosophically contentious about drawing this distinction, the relationship between two- and three- place trust becomes contested when one of these kinds of trust is taken to be, in some sense, more fundamental than the other. To be clear, it is uncontentious that philosophers, as Faulkner (2015: 242) notes, tend to 'focus' on three-place trust. What's contentious is whether—and if so, which—of these notions is theoretically more basic.

The overwhelming view in the literature maintains that three-place trust is the fundamental notion and that two-place (as well as one-place) trust are derivative upon three-place trust (e.g., Baier 1986; Holton 1994; Jones 1996; Faulkner 2007; Hieronymi 2008; Hawley 2014; cf., Faulkner 2015) Call this view *three-place fundamentalism*.

According to Baier, for instance, trust is centrally concerned with 'one person trusting another with some valued thing' (1986: 236) and for Hawley, trust is 'primarily a three-place relation, involving two people and a task' (2014: 2). We might think of two-place (e.g., X trusts Y) trust as derived from three-place trust in a way that is broadly analogous to how one might extract a diachronic view of someone on the basis of discrete interactions, as opposed to starting with any such diachronic view. On this way of thinking, three-place trust leads to two place trust over time, and is established on the basis of it.

Recent resistance to three-place fundamentalism has been advanced by Faulkner (2015) and Domenicucci and Holton (2017). Faulkner takes as a starting point that it is a desiderata on any plausible account of trust that it should accommodate infant trust, and thus, "that it not make essential to

trusting the use of concepts or abilities which a child cannot be reasonably believed to possess" (1986: 244). As Faulkner (2015) maintains, however, an infant, in trusting its mother 'need not have any further thought; the trust is no more than a confidence or faith – a trust, as we say – in his mother'. And so, Faulkner reasons, if we take Baier's constraint seriously, then we 'have to take two-place trust as basic rather than three-place trust.'

It is worth noting that this line of argument resembles a move that is highly contentious in other debates, such as in debates about whether knowledge is more theoretically basic than belief. Within this debate, it is contentious to move (as Jennifer Nagel 2013 has) from the fact that young children possess the concept of knowledge prior to possessing the concept of belief to the conclusion that it's false that belief is theoretically prior to knowledge (see McGlynn 2017). For the present purposes, it suffices to register that if ontogenetic arguments for theoretical primacy claims like Nagel's are contentious elsewhere, then plausibly, so will be Faulkner's attempt to move from the fact that infants can exhibit two-place trust prior to possessing the conceptual sophistication necessary for three-place trust to the conclusion that the latter is not theoretically more basic than the former.

A second strand of argument against three-place fundamentalism owes to Domenicucci and Holton (2017). According to Domenicucci and Holton, the kind of derivation of two-place trust from three-place trust that is put forward by three-place fundamentalists is implausible for other similar kinds of attitudes like love and friendship. As they put it:

> No one—or at least, hardly anyone—thinks that we should understand what it is for Antony to love Cleopatra in terms of the three place relation 'Antony loves Cleopatra for her __', or in terms of any other three-place relation. Likewise hardly anyone thinks that we should understand the two place relation of friendship in terms of some underlying three-place relation (here we don't even have any natural English expressions for the two-place). To this extent at least, we suggest that trust might be like love and friendship (2017: 149-50)

In response to this kind of argument by association, a proponent of three-place fundamentalism might either deny that these three- to two-place derivations are really problematic in the case of love or friendship, or instead grant that they are and maintain that trust is disanalogous.

In order to get a better sense of whether two-place trust might be unproblematically derived from three-place trust, regardless of whether the same holds *mutatis mutandis* for love in friendship, it will be helpful to look at a concrete attempt to do so. For example, according to Hawley (2014) three-place trust is analysed as: X relies on Y to phi because Y believes Y has a commitment to phi. And then, two-place trust defined simply as "reliance on someone to fulfil whatever commitments she may have" (2014: 16). If something like Hawley's reduction is unproblematic, then, as one line of response might go, this trumps whatever concerns one might have about the prospects of making analogous moves in the love and friendship cases.

c. Trust and belief: doxastic, non-doxastic and performance-theoretic accounts

Where does *belief* fit in to an account of trust? In particular, what beliefs (if any) must a truster have about whether the trustee will prove trustworthy? Proponents of *doxastic accounts* (e.g., Adler 1994; Hieronymi 2008; Keren 2014; McMyler 2011) hold that trust involves a *belief* on the part of the truster. On the simpler, straightforward incarnation of this view, when A trusts B to do X, A *believes* that B will do X. Other theorists propose more sophisticated belief-based accounts: on Hawley's (2019) account, for instance, to trust someone to do something is to believe that she has a commitment to doing it, and to rely upon her to meet that commitment. Conversely, to distrust someone to do something is to believe that she has a commitment to doing it, and yet not rely upon her to meet that commitment.

Non-doxastic accounts (e.g., Jones 1996; McLeod 2002; Paul Faulkner 2007; 2011; Baker 1987) have a negative and a positive thesis. The negative thesis is just the denial of the belief requirement on trust that proponents of doxastic accounts accept (viz., a denial that trusting someone to do something entails the corresponding belief that they will do that thing). This negative thesis, to note, is perfectly compatible with the idea that trust often times involves such a belief. What's maintained is that it is not essential. The positive thesis embraced by non-doxastic accounts involves a characterisation of some further non-doxastic attitude the truster, *qua* truster, must have with respect to the trustee's proving trustworthy.

An example of such a further (non-doxastic) attitude, on non-doxastic accounts, is *optimism*. For example, on Jones' (1996) view, you trust your neighbour to bring back the garden tools you loaned her only if you are optimistic that she will bring it back, and regardless of whether you believe she will. It should be pointed out that often times, optimism will lead to the acquisition of a corresponding belief. Importantly for Jones, the kind of optimism that characterises trust is not itself to be explained in terms of belief but rather in terms of affective attitudes entirely. Such a commitment is shared by non-doxastic views more generally which take trust to involve affective attitudes that might be apt to prompt corresponding beliefs.

Quite a few important debates about trust turn on the matter of whether a doxastic account or a non-doxastic accounts is correct. For example, discussions of the rationality of trust will look one way if trust essentially involves belief and another way if it does not (e.g., Jones 1996; Keren 2014). Relatedly, what one says about trust and belief will bear importantly on how one thinks about the relationship between trust and monitoring, as well as the distinction between paradigmatic trust and *therapeutic trust* (e.g., the kind of trust one engages in in order to build trust; see, e.g., Horsburgh 1960; Hieronymi 2008; Frost-Arnold 2014)

A notable advantage of the doxastic account is that it simplifies the epistemology of trust—and in particular, how trust can provide reasons for belief. Suppose, for example, that the doxastic account is correct, and so your trusting your colleague's word that they will return your laptop involves believing that they will return your laptop. This belief, some think, conjoined with the fact that your colleague tells you they will return your laptop, gives you a reason to believe that they will return your laptop. As Faulkner (2017: 113) puts it, on the doxastic account, '[t]rust gives a reason for belief because belief can provide reason for belief'. Non-doxastic accounts, by contrast, require further explanation as to why trusting someone would ever give you a reason to believe what they say.

Another advantage of doxastic accounts is that they are well-positioned to distinguish trusting someone to do something and mere optimistic wishing. Suppose, for instance, you loan £100 to a loved one with a terrible track record for repaying debts. Such a person may have lost your trust years ago, and yet you may remain optimistic and wishful that they will be trust worthy on this occasion. What distinguishes this attitude from genuine trust on the doxastic account is simply that you lack any belief that

your loved one will prove trustworthy. Explaining this difference is more difficult on non-doxastic accounts. This is especially the case on non-doxastic accounts according to which trust not only doesn't involve belief but positively precludes it, e.g., by essentially involving a kind of 'leap of faith' (Möllering 2006) that differs in important ways from belief.

Nonetheless, non-doxastic accounts have been emboldened in recent years in light of various serious objections that have been raised to doxastic accounts. One popular such objection highlights a key disanalogy with respect to how trust and belief interact with evidence, respectively. Here's Faulkner (2007):

> [T]rust need not satisfy either a positive or a negative evidence condition: it need not be based on evidence and can demonstrate a wilful insensitivity to the evidence. Indeed there is a tension between acting on trust and acting on evidence that is illustrated in the idea that one does not actually trust someone to do something if one only believes they will do it when one has evidence that they will. (2007: 876)

As Baker (1987) unpacks this idea, trusting can require ignoring counterevidence—as one might do when one trusts a friend despite evidence of guilt—whereas believing does not.

A particular type of example that puts pressure on doxastic accounts' ability to accommodate disanalogies with belief concerns *therapeutic trust*. In cases of therapeutic trust, the purpose of trusting is to *promote* trustworthiness and is not predicated on prior belief of trustworthiness. Take again the case of loaning £100 to a loved one with a terrible track record for repaying debts. It might be that when one makes this loan one has given up entirely on the loved one and no expectations that the loan will be repaid or that the trustee's behaviour will change. In such a case, it looks implausible that trust is present.

However, suppose we flesh out the case further. Suppose you trust with the aim of *establishing* trust (e.g., Frost-Arnold 2014; Faulkner 2011), as one might trust a teenager with an important task, in hopes that by trusting them, it will then lead them to become more trustworthy in the future. In this kind of case, we are plausibly trusting, but not on the basis of prior evidence or belief we have that the trustee will succeed on this occasion. To the extent that

such a description of this kind of case is right, therapeutic trust offers a counterexample to the doxastic account, as it involves trust in the absence of belief.

A third kind of account—the *performance-theoretic* account of trust (e.g., Carter 2019)—makes no essential commitment as to whether trusting involves belief. On the performance-theoretic account, what's essential to the attitude of trusting is how it is *normatively constrained*. An attitude is a trust attitude (toward a trustee, T, and a task, X) just in case the attitude is successful if and only if the T takes care of X as entrusted. Just as there is a sense in which, for example, your archery shot is not successful if it misses the target (see, e.g., Sosa 2010; 2015; Carter forthcoming), your trusting someone to keep a secret misses its mark, and so fails to be successful trust, if the trustee spills the beans. With reference to this criterion of successful (and unsuccessful) trust, the performance-theoretic account aims to explain what good and bad trusting involves (see §2.a), and also why some trust is more valuable than others (see §3).

d. Deception detection and monitoring

Given that trusting inherently involves the incurring of some level of risk to the trustee, it is natural to think that trust would in some way be *improved* by the truster doing what she can to minimize such risk, e.g., by monitoring the trustee with an eye to, e.g., preempting any potential betrayal or at least mitigating the expected disvalue of potential betrayal.

This *prima facie* plausible suggestion, however, raises some perplexities. As Annette Baier (1986) puts it:

> Trust is a fragile plant [...] which may not endure inspection of its roots, even when they were, before inspection, quite healthy (1986: 260)

There is something intuitive about this point. If, for instance, A trusts B to drive the car straight home after work--but then proceeds to surreptitiously drive behind A the entire way in order to make sure that B really does drive

straight home, it seems that A in doing so *no longer* is trusting B. The trust, it seems, dissolves through the process of such monitoring.

Extrapolating from such cases, it seems that trust *inherently* involves not only subjecting oneself to some risk, but also *remaining* subjected to such risk--or, at least--behaving ways that are compatible with one's *viewing* oneself as (remaining to be) subjected to such risk.

The above idea of course needs sharpened. For example, trusting is plausibly not destroyed by *negligible* monitoring, or perhaps of monitoring of any sort whatsoever. The crux of the idea seems to be, as Faulkner (2011, §5) puts it, that *'too much* reflection' on the trust relation, perhaps in conjunction with making attempts to minimise risks that trust will be betrayed, can undermine trust. A specification of what 'too much' involves, however, remains a difficult question.

One form of monitoring—construed loosely—that is plausibly compatible with trusting is *contingency planning* (e.g., Carter 2020). For example, suppose you trust your teenager to drive your car to work and back in order that they may undertake a summer job. A prudent mitigation against the additional risk incurred (e.g., that the car will be wrecked in the process) will be to buy some additional insurance upon trusting the car with the teenager. The purchasing of this insurance, however, doesn't itself undermine the trusting relationship, even though it involves a kind of risk mitigating behaviour.

One explanation here turns on the distinction between (i) mitigating against the risk that trust will be betrayed; and (ii) mitigating against the extent or severity of the harm or damage incurred *if* trust is betrayed. Contingency planning involves type-(ii) mitigation, whereas, for example, trailing behind the teenager with your own car, which is plausibly incompatible with trusting, is of type-(i).

2. The normativity of trust

Norms of trust arise between the two parties of reciprocal trust: a norm to be trusting in response to the invitation to trust, and to be trustworthy in response to the other's trusting reliance (e.g., Fricker 2018). The former normativity lies 'on the truster's side', and the latter on the trustee's side. In

this section, we discuss norms on trusting by looking at these two kinds of norms—that govern the truster and the trustee, respectively—in turn.

a. On the truster's side

This section will discuss, first, (i) *general* norms on trusting on the truster's side, and then will engage—in some detail—with the complex issue of the norms governing trust *in another's words* specifically.

(i) *Entitlement to trust, obligation to trust*

If—as doxastic accounts maintain—trust is a species of belief (Hieronymi 2008), then the rational norms governing trust govern belief, such that (for example) it will be irrational to trust someone whom you have strong evidence is unreliable, and the norm violation here is the same kind of norm violation in play in a case where one simply believes, against the evidence, that that individual is trustworthy. Thus: to the extent that one is rationally entitled to *believe* the trustee is trustworthy, with respect to Φ one thereby has an entitlement (on these kinds of views) to trust the trustee to Φ.

The norms that govern trust on the truster's side will look different on non-doxastic accounts. For example, on a proposal like Frost-Arnold's (2014), according to which trust is characterised as a kind of non-doxastic *acceptance* rather than as belief, the rationality governing trusting will be the rationality of *acceptance*, where rational acceptance can in principle come apart from rational belief. For one thing, whereas the rationality of belief is exclusively epistemically constrained, the rationality of acceptance needn't be. In cases of therapeutic trust, for example, it might be *practically* rational (viz., rational with reference to the adopted end of building a trusting relationship) to accept that the trustee will Φ, and thus, to use the proposition that they will Φ as a premise in practical deliberation (see, e.g., Bratman 1992; Cohen 1989)—i.e., to act as if it is true that they will Φ. Of course, acting *as if* a proposition is true neither implies nor is implied by believing that it is true.

On performance-theoretic accounts, trusting is subject, on the truster's side, to three kinds of *evaluative* norms, which correspond with three kinds of positive evaluative assessments: *success, competence*, and *aptness*. Whereas trusting is successful if and only if the trustee takes care of

things as intrusted, trusting is *competent* if and only if one's trusting issues from a reliable disposition—viz., a competence—to trust successfully when appropriately situated (for discussion, see Carter 2019).

Just as successful trust might be *incompetent*, e.g., as when one trusts someone with a well-known track record of unreliability who happens to prove trustworthy on this particular occasion, likewise, trust might fail to be successful *despite* being competent, e.g., as when one trusts an ordinarily reliable individual who, due to fluke luck, fails to take care of things as entrusted on this particular occasion. Even if trust is both successful *and* competent, however, there remains a sense in which it could fall short of the third kind of evaluative standard—viz., *aptness*. Aptness demands success *because* competence, and not merely success *and* competence (see, e.g., Sosa 2010; 2015; Carter 2019; forthcoming). Trust is apt, accordingly, if and only if one's trusts successfully such that the successful trust manifests her trusting competence.

A final species of norm that merits discussion on the truster's side is a norm of *obligation*. Obligations to trust can be generated, trivially, by promise-making (cf., Owens 2017) or by other kinds of cooperative agreements (Faulkner 2011, Ch. 1). Of more philosophical interest are cases where obligations to trust are generated outwith explicit agreements.

One case of particular interest here arises in the literature on *testimonial injustice*, pioneered by Miranda Fricker (2007). Put roughly, testimonial injustice occurs when a speaker receives an unfair deficit of credibility from a hearer due to prejudice on the hearer's part, resulting in the speaker's being prevented from sharing what she knows.

An example of testimonial injustice that Fricker uses as a reference point is from Harper Lee's *To Kill a Mockingbird*, where Tom Robinson, a black man on trial after being falsely accused of raping a white woman, has his testimony dismissed due to the prejudiced preconceptions on the part of the jury which owes to deeply seated racial stereotypes. In this case, the jury has makes a deflated credibility judgement of Robinson, and as a result, he is unable to convey to them the knowledge that he has of the true events which occurred.

On one way of thinking about norms of trusting on the truster's side, the members of the jury have mere *entitlements* to trust Robinson's testimony though no obligation to do so; thus, their distrust of Robinson is not norm-violating. This gloss of the situation, on Fricker's view, is

incomplete (see also, e.g., Medina 2011; 2013); it fails to take into account the sense in which Robinson is *wronged* as a result of this distrust. An appreciation of this wrong, according to Fricker, Medina, and others who countenance epistemic injustice as a serious epistemic (as well as moral) harm, should lead us to think of the relevant norm on the hearer's side as a norm of *obligation*; as such, on this view, *distrust* that arises from affording a speaker a prejudiced credibility deficit is not merely an instance of foregoing trusting when one is entitled to trust, but failing to trust when one should. For additional work discussing the relationship between trust and testimonial injustice, see, e.g., Origgi (2012); Medina (2011); Piovarchy (forthcoming); Pohlhaus Jr (2014); Wanderer (2017); Carter and Meehan (Forthcoming).

(ii) *Trust in Words*

Why not *lie*? (Or, more generally, why not: promise to take care of things, and then renege on that promise whenever it is convenient to do so?) According to a fairly popular answer (Faulkner 2011; Simion 2020), deception is bad not only for the deceived, but it is bad likewise for the *deceiver* (see also, Kant). If one cultivates a reputation as being untrustworthy, then this comes with *practical* costs in social communities; the untrustworthy person, *recognised as such*, is outcast, and *de facto* foregoes the (otherwise possible) social benefits of trusting.

Things are more complicated, however, in *one-off* trust-exchanges—where the risk of the disvalue of cultivating an untrustworthy reputation is minimal. We may repose the question within the one-off context: why not lie and deceive, when it is convenient to do so, in one-off exchanges? In one-off interactions where we (i) don't know others' motivations but (ii) do appreciate that there's a *general* motivation to be unreliable (e.g., to reap gains of betrayal), it is surprising that we find as much trustworthy behaviour as we do. Why don't people betray to a greater extent than they do in such circumstances, given that betrayal seems *prima facie* the most rational decision-theoretic move?

According to Faulkner, when we communicate with another as to the facts, we face a situation akin to a *prisoner's dilemma* (2011: 6). In a prisoner's dilemma, our aggregate well-being will be maximized if we both cooperate. However, given the logic of the situation, it looks like the rational

thing to do for each of us is to defect. We're then faced with a problem: how to ensure the cooperative outcome?

Similarly, Faulkner argues, speakers and audiences have different interests in communication. The audience is interested in learning the truth. I contrast, engaging in conversations is to the advantage of speakers because it is a means of influencing others: through an audience's acceptance of what we say, we can get an audience to think, feel, and act in specific ways. So, according to Faulkner, our interest, qua speaker, is being believed, because we have a more basic interest in influencing others. The commitment to telling the truth would not be best for the speaker. The best outcome for a speaker would be to receive an audience's trust and yet have the liberty to tell the truth or not. (2011: 5-6).

There are four main reactions to this problem in the literature. According to Reductionism in the epistemology of testimony (Adler 1994; Audi 1997; 2004; 2006; Faulkner 2011; Fricker 1994; 1995; 2017; 2018; Hume 1739; Lipton 1998; Lyons 1997), in virtue of this lack of alignment of hearer and speaker interests, one needs positive, independent reasons to trust their speaker: since communication is like a prisoner's dilemma, the hearer needs a reason for thinking or presuming that the speaker has chosen the cooperative, helpful outcome. Anti-Reductionism (e.g., Burge 1993; 1997; Coady 1973; 1992; Goldberg 2006; 2010; Goldman 1999; Graham 2010; 2012a; 2015; Greco 2019; 2015; Green 2008; Reid 1764; Simion 2020; Simion and Kelp 2018) rejects this claim. According to these philosophers, we have a default entitlement to believe what we are being told. In turn, this default entitlement is derivable a priori from the nature of reason (e.g., Burge 1993; 1997) sourced in social norms of truth-telling (Graham 2012b) social roles (Greco 2015) the reliance on other people's justification-conferring processes (Goldberg 2010) or the knowledge norm of assertion (Simion 2020). Other than these two main views, we also encounter hybrid views (Lackey 2003; 2008; Pritchard 2004) that try to impose weaker conditions on testimonial justification than Reductionism, while, at the same time, not being as liberal about it as Anti-Reductionism.

Another reaction to Faulkner's problem of cooperation for testimonial exchanges is scepticism (e.g., Graham 2012a; Simion 2020); for the sceptic, it is not even clear that the problem gets off the ground to begin with, and if it does, it's not clear whether it's a problem rather than a mere challenge to identify the source of de facto speaker reliability. Recall that according to

Faulkner, in testimonial exchanges, the default position for speakers involves no commitment to telling the truth. If that is the case, he argues, the default position for hearers involves no entitlement to believe.

It is important to understand the significance of the '*default*' position at stake in the argument. What is at stake in the debate between Reductionism and Anti-reductionism is prima facie entitlement. Both views will allow, for instance, that if there are defeaters present at the relevant context, hearers are not entitled to believe. This explains the concern with what the default, or starting position is in testimonial exchanges: all else absent, are we entitled to believe what we are being told? Anti-Reductionism says 'yes'; Reductionism says 'no'—viz., you need to go out in the world and look for further support for your belief.

Since the Problem of Cooperation is supposed to be a first-personal decision-theoretic problem, modelled on the Prisoners' Dilemma, it seems that it needs to be formulated accordingly, as concerning a problem of first-personal perspective. The problem then would take something like the following shape: when the audience knows nothing else about the speaker, the audience still knows that it is in a speaker's interest not to be constrained in their testimony. The later, in turn, is a reason against believing what the speakers says. If you have a reason to believe that a speaker's interest in the communicative exchange is getting you to believe what they say whether it's true or not, and you have no other information about the speaker, this reason is in itself a reason to not believe what is said. This, in turn, requires a 'defeater defeater' for entitlement, which is why there is a standing demand that the audience have some positive reason for testimonial uptake.

Here is what this version of the argument looks like:

> (1) Hearers know that they are interested in truth and that speakers are interested in being believed.
> (2) Hearers know that the *default* position for speakers is seeing to their own interests rather than to the interests of the hearers.
> (3) Therefore, hearers know that it is not the case that the default position for speakers is telling the truth (from 1 and 2).
> (4) The default position for hearers is trust only if they don't know that it is not the case that the default position for speakers is telling the truth.

(5) Therefore it is not the case that the default position for hearers is trust (from 3 and 4)

Note, though that this reading departs significantly from the 'default position' that constitutes the beef between Reductionism and Anti-Reductionism to begin with: the hearer, on this reading of the problem of cooperation, has quite a bit of information against believing the speaker's words, at least in the absence of further information. Just like in the cases described in the empirical studies on contextually-informed deception detection (e.g., Kraut 1980; Bond Jr. and DePaulo 2006; deTurck et al. 1990) the hearer in this case is in possession of antecedent information undercutting the testimonial source in virtue of speaking against its credentials. There is a question, then, whether this first-personal reading of the problem of cooperation is a genuine way of depicting the debate between Reductionism and Anti-Reductionism: again, reductionists and anti-reductionists alike will agree that, in the present of defeat, the hearer lacks entitlement to believe (Simion 2020).

One way to respond is to claim that this *is* the default position in testimonial exchanges, in that *all cases* of testimony exhibit defeat in this way. As such, in *all* cases of testimony it is the case that one needs positive reasons to believe one's speaker: that is the default position. Two things about this reply, though: first, if this is right, the disagreement between Reductionism and Anti-Reductionism disappears. Since the two views agree that, in cases of defeat, defeater defeaters are needed for entitlement, and since this reply takes the default position to be one in which defeat is present, the debate over entitlement in the default position vanishes.

Second, there are two further, independent worries concerning the plausibility of premises (1) to (3). First, it is not clear that this is the correct utility profile of the case: are *all* speakers really such that they care about being believed? This seems like a fairly heavy empirical assumption. More importantly, though, are all hearers such that they have all the knowledge ascribed to them in (1) to (3)? After all, these premises assign a whole lot of quite sophisticated knowledge to your everyday hearer. Even if Faulkner is right about the utility profile of the case, the assumption that everyone who ever engages in testimonial exchanges knows that this is the utility profile is certainly quite implausible, since many years of philosophical theorizing have been put into assuming the conjunction of (1), (2) and (3). If all of this is right, though, the first-personal reading of the problem of cooperation will not do

the needed work in the Reductionism-Anti-Reductionism debate (Simion 2020; Simion and Kelp 2018).

Maybe, then, what is needed here is a reading of the problem of cooperation that is third-personal, in that it does not presuppose any knowledge on the part of the hearer. Here is how such an argument would go:

> (1) Hearers are interested in truth; speakers are interested in being believed.
> (2) The *default* position for speakers is seeing to their own interests rather than to the interests of the hearers.
> (3) Therefore, it is not the case that the default position for speakers is telling the truth (from 1 and 2).
> (4) The default position for hearers is trust only if the default position for speakers is telling the truth.
> (5) Therefore it is not the case that the default position for hearers is trust (from 3 and 4) .

This way of looking at the problem, indeed, does not suffer from any of the drawbacks identified for the first-personal reading, in virtue of not presupposing any knowledge on the hearer's side. Here is the main worry with this reading, though: on the reconstruction above, the conclusion does not follow. In particular, the problem is with premise (3), which is not supported by (1) and (2) (Simion 2020). That is because being interested in being believed does not exclude also being interested in telling the truth. Speakers might still—by default—also be interested in telling the truth, on independent grounds, that is, independently of their concern (or, rather, lack thereof) with hearer's interests; indeed, the sources of entitlement proposed by the Anti-Reductionist—e.g. the existence of social norms of truth telling, the knowledge norm of assertion etc.—may well constitute themselves in reasons for the speaker to tell the truth—absent overriding incentive to do otherwise. If that is the case, telling the truth will be default for hearers, therefore trust will be default for hearers.

According to Faulkner himself, trust lies at the heart of the solution to his problem of cooperation, i.e., it gives the speakers reasons to tell the truth (2011, Ch. 1; 2017). Faulkner thinks that the problem of cooperation that affects testimonial exchanges is resolved 'once one recognizes how *trust itself* can give reasons for cooperating' (2017: 9)

According to Faulkner, when the hearer H believes that speaker S can see that H is relying on S for information about whether p, and in addition H trusts S for that information, then H will make a number of presumptions: 1. H believes that S recognizes H's trusting dependence on S proving informative. 2. H presumes that if S recognizes H's trusting dependence, then S will recognize that H normatively expects S to prove informative. 3. H presumes that if S recognizes H's expectation that S should prove informative, then other things being equal, S will prove informative for this reason. 4. So taking the attitude of trust involves presuming that the trusted will prove trustworthy (2011: 130). The hearer's presumption that the speaker will prove informative rationalizes the hearer's uptake of the speaker testimony.

Furthermore, Faulkner claims, H's trust gives S "a reason to be trustworthy," such that S is, as a result, more likely to be trustworthy: it raises the objective probability that S will prove informative in utterance. In this fashion, "acts of trust can create as well as sustain trusting relations" (2011: 156-7). As Graham (2012a) puts it, "the hearer's trust—the hearer's normative expectation, which rationalizes uptake—then "engages," so to speak, the speaker's internalization of the norm, which thereby motivates the speaker to choose the informative outcome." Speakers who have internalized these norms will then often enough choose the informative outcome when they see that audiences need information; they will be "motivated to conform" because they have "internalized the norm" and so "intrinsically value" compliance (2011: 186). As such, the de facto reliability of testimony is explained by the fact that the trust placed in hearers by the speakers triggers, on the speakers' side, the internalization of social norms of trust, which, in turn, makes speakers objectively likely to put hearers' informational interests before their own.

One worry for Faulkner's picture is that his own solution threatens, one more time, to dissolve the problem of cooperation rather than solve it (Graham 2012a). Recall how the problem was set-up: the thought was that speakers only care about being believed, whether they are speaking the truth or not, which is why the hearer needs some reason for thinking the speaker is telling them the truth. But if speakers have internalized social norms of trustworthiness, it's not true that speakers are just as apt to prove uninformative as informative. It's not true that they're only interested in being believed. Rather they are out to inform, to prove helpful; due to having

internalised the relevant trustworthiness norms, speakers are committed to informative outcomes (Graham 2012a).

b. On the trustee's side

We have seen that trust can be a two-place or a three-place relation. In the former case, it is a relation between a trustor and a trustee, as in Ann trusts George. Two-place trust seems to be a fairly highbrow affair: when we say that Ann trusts George *simpliciter*, we seem to attribute a fairly robust attitude to Ann, whereby she trusts him in (at least) several respects. In contrast, three-place trust is a less involved affair: when we say that Ann trusts George to do the dishes, we need not say much about their relationship otherwise.

This contrast is preserved when we switch from focusing on the trustor's trust to the trustee's trustworthiness. That is, one can be trustworthy *simpliciter* (corresponding to a two-place trust relation) but one can also be trustworthy with regard to a particular matter—i.e., two-place trustworthiness (Jones 1996) corresponding to three-place trust. For instance, my surgeon might well be extremely trustworthy when it comes to performing surgery well, but not in any other respects.

Some people working on trustworthiness focus more on two-place trust. As such, since the two-place trust relation is surely a more robust one, they put forward accounts of trustworthiness that are generally quite demanding, less trivially met, in that they require the trustee to be reliably making good on their commitments, but also to do so out of the right motive.

The classic such account is Annette Baier's *goodwill* based account (see also e.g., Potter 2002); in a similar vein, others combine reliance on goodwill with certain expectations (Jones 1996) including in one case a normative expectation of goodwill (Cogley 2012). According to this kind of view, the trustworthy person fulfils their commitments *in virtue of* their goodwill towards the trustor. This view, according to Baier, makes sense of the intuition that there is a difference between trustworthiness and mere reliability, that corresponds to the difference between trust and mere reliance.

The most widely spread worry about these accounts of trustworthiness is that they are too strong: we can trust other people without presuming that they have goodwill. Indeed, our everyday trust in strangers falls into this

category. If so, the argument goes, this seems to suggest that whether or not people are making good on their commitments out of goodwill or not is largely inconsequential: "[w]e are often content to trust without knowing much about the psychology of the one-trusted, supposing merely that they have psychological traits sufficient to get the job done" (Blackburn 1998).

Another worry for these accounts is that, while more plausible as accounts of trustworthiness *simpliciter*, they give quite counterintuitive results in the case of two-place trustworthiness: indeed, whether George is trustworthy when it comes to washing the dishes or not seems to not depend on his goodwill, nor on other such noble motives. The goodwill view is too strong.

Unfortunately, it looks as though there is reason to believe the goodwill view is, at the same time, too weak. To see this, consider the case of a convicted felon and his mother: it looks as though they can have a goodwill-based relationship, and thus be trustworthy within the scope thereof, while, at the same time, not being someone whom we would describe as trustworthy (Potter 2002: 8).

If all of this is the case, it begins to look as though the presence of goodwill is independent of the presence of trustworthiness. This observation motivates accounts of trustworthiness that rely on less highbrow motives underlying the trustee's reliability. One such account is the social contract view of trustworthiness. According to this view, the motives underlying people's making good on their commitments are sourced in social norms and the unfortunate consequences to one's reputation and general wellbeing of breaking them (Hardin 2002: 53; see also O'Neill 2002; Dasgupta 2000). Self-interest determines trustworthiness on these accounts.

It is easy to see that social contract views do well in accounting for trustworthiness in three-place-trust relations: George is trustworthy when it comes to washing the dishes, on this view: he makes good on his commitments in virtue of social norms making it such that it's in his best interest to do so. The main worry for these views, however, is that they will be too permissive, and thus have difficulties in distinguishing between trustworthiness proper and mere reliability. Relatedly, the worry goes, these views seem less well equipped to deal with trustworthiness simplicter, i.e. corresponding to a two-place trust relation. For instance, on a social contract view, it would seem that a sexist employer who treats female employees well only because he believes that he would face legal sanctions if he did not, will

come out as trustworthy (Potter 2002, 5). This is an intuitively unfortunate result.

One thought that gets prompted by the case of the sexist employer is that trustworthiness is a character trait that virtuous people possess; after all, this seems to be something that the sexist employer is missing. On Nancy Potter's view, trustworthiness is a disposition to respond to trust in appropriate ways, given "who one is in relation" to and given other virtues that one possesses or ought to possess (e.g., justice, compassion) (2002: 25). According to Potter, a trustworthy person is "*one who can be counted on, as a matter of the sort of person he or she is, to take care of those things that others entrust to one.*"

When it comes demandingness, the virtue-based view seems to lie somewhere in-between the goodwill view, on one hand, and the social contract view, on the other. It seems more permissive than the former in that it can account for the trustworthiness of strangers insofar as they display the virtue .at stake. It seems more demanding than the latter in that it purports to account for the intuition that mere reliability is not enough for trustworthiness: rather, what is required is reliability sourced in good character.

Recent criticism of virtue-based views comes from Jones (2012). According to her, trustworthiness does not fit the normative profile of virtue, in the following way: if trustworthiness were a virtue, then being untrustworthy would be a vice. However, according to Jones, that cannot be right: after all, we are often required to be untrustworthy in one respect or another – for instance, because of conflicting normative constraints–but it cannot be that being *vicious* is ever required.

Other serious problems for Potter's specific view are its apparent un-informativeness; first, defining the trustworthy person as 'a person who can be counted on as a matter of the sort of person he or she is' threatens vicious circularity: after all, it defines the trustworthy as those that can be trusted. Relatedly, the account turns out to too vague to give definite predictions in a series of cases. Take again the case of the sexist employer: why is it that he cannot be 'counted on, as a matter of the sort of person he or she is, to take care of those things that others entrust to one' in his relationship with his female employees? After all, in virtue of the sort of person he is – i.e., the sort of person who cares about not suffering the social consequences of mistreating them – he can be counted on to treat his employees well. If that

is so, Potter's view will not do much better than social contract views when it comes to distinguishing trustworthiness from mere reliability.

More recently, several philosophers propose purely externalist accounts of trustworthiness. Katherine Hawley's (2019) view falls into this camp. According to her, trustworthiness is a matter of avoiding unfulfilled commitments, which requires both caution in incurring new commitments and diligence in fulfilling existing commitments. Crucially, on this view, one can be trustworthy regardless of one's motives for fulfilling one's commitments. Hawley's is a negative account of trustworthiness, which means that one can be trustworthy whilst avoiding commitments as far as possible. Untrustworthiness can arise from insincerity or bad intentions, but it can also arise from enthusiasm and becoming over-committed. A trustworthy person must not allow her commitments to outstrip her competence.

One natural question that arises for this view is: how about commitments that we do not, but we should take on board? Am I a trustworthy friend if I never take on any commitments towards my friends? According to Hawley, in practice, through friendship, work and other social engagements we take on meta-commitments—commitments to incur future commitments. These can make it a matter of trustworthiness to take on certain new commitments.

Another view in a similar, externalist vein is developed by Simion and Kelp (2020b). According to them, trustworthiness is a disposition to fulfil one's obligations. What drives the view is the thought that one can fail to fulfil one's commitments in virtue of being in a bad environment–an environment that 'masks' the normative disposition in question—while, at the same time, remaining a trustworthy person. Again, on this view as well, whether the disposition in question is there in virtue of good will or not is inconsequential. That being said, Simion and Kelp's view can accommodate the thought that people who comply with a particular norm for the wrong reason are less trustworthy than their goodwilling counterparts. To see how, take the sexist employer again: insofar as it is plausible that there are norms against sexism, as well as norms against mistreating one's female employees, the sexist employer fulfils the obligations generated by the latter but not by the former. In this, he is trustworthy when it comes to treating his employees well, but not trustworthy when it comes to treating them well for the right reason. As such, the Simion and Kelp's view does go in the direction of accounting for

the intuitions of those pressing the difference between trustworthiness and mere reliability in fulfilling a particular commitment.

Another advantage for the view is that it explains the intuitive difference in robustness between two-place trustworthiness and trustworthiness simpliciter. According to this view, one is trustworthy simpliciter when one meets a contextually-variant threshold of two-place-trustworthiness for contextually salient-obligations. For instance, a philosophy professor is trustworthy simpliciter in the philosophy department just in case she has a disposition to meet enough of her contextually salient obligations: do her research and teaching, not be late for meetings, answer emails promptly, help students with their essays etc. The same philosophy professor, however, will be trustworthy at home just in case she e.g. helps the kids with their homework, doesn't neglect her partner etc.

3. The Value of Trust

Trust is *valuable*. Without it, we face not only cooperation problems, but we also incur substantial risks to our well-being—viz., those ubiquitous risks to life that characterise—at the limit case—the Hobbesian (1651/1970) 'state of nature'. Accordingly, one very general argument for the value of trust appeals to the disutility of its absence (see also Alfano forthcoming).

Moreover, apart from merely serving as an enabling condition for other valuable things (e.g., the possibility of large-scale collective projects for societal benefit), trust is also instrumentally valuable for both the truster and the trustee as a way of resolving particular (including one-off) cooperation problems in such a way as to facilitate mutual profit (see §2). Furthermore, trust is instrumentally valuable as a way of *building* trusting relationships (e.g., Solomon and Flores 2003). For example, trust can effectively be *used*—as when one trusts a teenager with a car to help cultivate a trust relationship—in order to make more likely the attainments of the benefits of trust (for both truster and trustee) further down the road (e.g., Horsburgh 1960; Jones 2004; Frost-Arnold 2014).

Apart from the trust's uncontroversial *instrumental* value (for helpful discussion, see O'Neill 2002), trust can be, more controversially, *finally* valuable, at least in certain circumstances. Something X is instrumentally

valuable, with respect to an end, Y in so far as it is valuable *a means to* Y; instrumental value can be contrasted with final value. Something is finally valuable iff it is valuable *for its own sake*. An example of something instrumentally valuable is money, which we value because of its usefulness in getting other things; an example of something (arguably) finally valuable is beauty or pleasure.

One way to defend the view that trust can—at least in certain circumstances—be finally valuable, and not merely instrumentally valuable, is to supplement the performance-theoretic view of trust (see §1.c and §2.a) with some additional (albeit somewhat contentious) axiological premises as follows:

(P1) Apt trust is successful trust that is because of trust-relevant competence (From the performance-theoretic view of trust)
(P2) Achievements are successes that are because of competence (Premise)
(C1) So, apt trust is an achievement (from P1 and P2)
(P3) Achievements are finally valuable (Premise)
(C2) So, apt trust has final value (from C1 and P3)

Premise (2) of the argument—which defines achievement as 'success because of ability' is mostly uncontentious; at least, the identification of achievement as a normative kind with the 'success from ability' schema is taken for granted widely in contemporary virtue epistemology (e.g., Greco 2010; 2009; Haddock, Millar, and Pritchard 2009; Sosa 2010b) and elsewhere (e.g., Feinberg 1970; Bradford 2013; 2015).

Premise (3) however is where the action lies. Even if apt trust is an achievement given that it involves a kind of success because of ability (i.e., trust-relevant competences), we'd need some positive reason to connect the 'success because of ability' structure with final value if we are to accept (P3).

One such argument for making this type of connection—viz., a connection between a success's being finally valuable and facts about how the success came about—is found in the value theory literature (e.g., Rabinowicz and Ronnow-Rasmussen 2000; Rønnow-Rasmussen 2003). According to Rabinowicz and Ronnow-Rasmussen, some things are valuable for their own sake on account of their *intrinsic* properties. But other things which are valuable for their own sake are valuable for their own sake on account of their

non-intrinsic, *relational* properties. These latter kinds of things (consider, for example, the value of the first book off of the Gutenberg printing press, a dress worn by Lady Diana, etc.) are finally valuable in that their value is not exhausted by the value they have as a means to an end. At the same time, though, such things are not intrinsically valuable. And this is because the *supervenience base* of the final value they have is not exhausted by their intrinsic properties (otherwise, intrinsic duplicates would have equal value), and includes at least partly their relational properties, e.g., the connection they bear to the Gutenberg press, to Lady Diana, etc.

What this all suggests is that the matter of whether (3) is true doesn't turn on the matter of whether achievements are intrinsically valuable, but rather, on whether they are finally valuable in virtue of the connection achieved successes have to achievements themselves.

Are achievements, then, axiologically akin to books off the first printing press, dresses work by Diana, etc., or not? A strong line here defends (3) by maintaining that *all* achievements (including evil achievements and 'trivial' achievements) are finally valuable, because successes because of ability (no matter what the success, no matter what the ability used) have a value that is not reducible to just the value of the success.

This kind of argument faces some well-worn objections (for some recent helpful discvussions, see, Kelp and Simion 2016; Dutant 2013; Goldman and Olsson 2009; Sylvan 2017). A more nuanced line of argument for C2 will weaken (3) so that it says, instead, that (3*) *some* achievements are finally valuable. But with this weaker premise in play, (3*) and (C1) no longer entail C2; what would be needed—and this remains an open problem for work on the axiology of trust—is a further premise to the effect that the kind of achievement that features in *apt trust,* specifically, is among the finally valuable rather than non-finally valuable achievements. And a defence of such a further premise, of course, will turn on further considerations about (among other things) the value of successful and competent trust, perhaps also in the context of wider communities of trust.

References

Adler, Jonathan E. 1994. 'Testimony, Trust, Knowing'. *The Journal of Philosophy* 91 (5): 264–275.

Alfano, Mark. forthcoming. 'The Topology of Communities of Trust'. *Russian Sociological Review*.

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

———. 2004. 'The a Priori Authority of Testimony'. *Philosophical Issues* 14: 18–34.

———. 2006. 'Testimony, Credulity, and Veracity'. In *The Epistemology of Testimony*, edited by Jennifer Lackey and Ernest Sosa, 25–49. Oxford University Press.

Baier, Annette. 1986. 'Trust and Antitrust'. *Ethics* 96 (2): 231–260. https://doi.org/10.1086/292745.

Baker, Judith. 1987. 'Trust and Rationality'. *Pacific Philosophical Quarterly* 68 (1): 1–13. https://doi.org/10.1111/j.1468-0114.1987.tb00280.x.

Blackburn, Simon. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford University Press UK.

Bond Jr, Charles F., and Bella M. DePaulo. 2006. 'Accuracy of Deception Judgments'. *Personality and Social Psychology Review* 10 (3): 214–234.

Bradford, Gwen. 2013. 'The Value of Achievements'. *Pacific Philosophical Quarterly* 94 (2): 204–224.

———. 2015. *Achievement*. Oxford University Press.

Bratman, M. 1992. 'Practical Reasoning and Acceptance in a Context'. *Mind* 101 (401): 1–16.

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

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

Carter, J. Adam. forthcoming. 'De Minimis Normativism: A New Theory of Full Aptness'. *Philosophical Quarterly*.

———. 2019. 'On Behalf of a Bi-Level Account of Trust'. *Philosophical Studies*, 1–24.

———. 2020. 'Therapeutic Trust'. *Manuscript*.

Carter, J. Adam, and Daniella Meehan. Forthcoming. 'Trust, Distrust, and Epistemic Injustice'. *Educational Philosophy and Theory*.

Coady, C. A. J. 1973. 'Testimony and Observation'. *American Philosophical Quarterly* 108 (2): 149–55.

———. 1992. *Testimony: A Philosophical Study*. Oxford University Press.

Cogley, Zac. 2012. 'Trust and the Trickster Problem'. *Analytic Philosophy* 53 (1): 30–47. https://doi.org/10.1111/j.2153-960X.2012.00546.x.
Cohen, L. Jonathan. 1989. 'Belief and Acceptance'. *Mind* 98 (391): 367–389.
Dasgupta, Partha. 2000. 'Trust as a Commodity'. *Trust: Making and Breaking Cooperative Relations* 4: 49–72.
deTurck, Mark A, Janet J Harszlak, Darlene J Bodhorn, and Lynne A Texter. 1990. 'The Effects of Training Social Perceivers to Detect Deception from Behavioral Cues'. *Communication Quarterly* 38 (2): 189–199.
Domenicucci, Jacopo, and Richard Holton. 2017. 'Trust as a Two-Place Relation'. *The Philosophy of Trust*, 149–160.
Dutant, Julien. 2013. 'In Defence of Swamping'. *Thought: A Journal of Philosophy* 2 (4): 357–366.
Faulkner, Paul. 2007. 'A Genealogy of Trust'. *Episteme* 4 (3): 305–321. https://doi.org/10.3366/E174236000700010X.
———. 2011. *Knowledge on Trust*. Oxford: Oxford University Press.
———. 2015. 'The Attitude of Trust Is Basic'. *Analysis* 75 (3): 424–429.
———. 2017. 'The Problem of Trust'. *The Philosophy of Trust*, 109–28.
Feinberg, Joel. 1970. *Doing and Deserving; Essays in the Theory of Responsibility*. Princeton: Princeton University Press.
Fricker, Elizabeth. 1994. 'Against Gullibility'. In *Knowing from Words*, 125–161. Springer.
———. 1995. 'Critical Notice'. *Mind* 104 (414): 393–411.
———. 2017. 'Inference to the Best Explanation and the Receipt of Testimony: Testimonial Reductionism Vindicated'. *Best Explanations: New Essays on Inference to the Best Explanation*, 262–94.
———. 2018. *Trust and Testimonial Justification*.
Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press Oxford.
Frost-Arnold, Karen. 2014. 'The Cognitive Attitude of Rational Trust'. *Synthese* 191 (9): 1957–1974.
Goldberg, Sanford C. 2006. 'Reductionism and the Distinctiveness of Testimonial Knowledge'. *The Epistemology of Testimony*, 127–44.
———. 2010. *Relying on Others: An Essay in Epistemology*. Oxford University Press.
Goldman, Alvin I. 1999. 'Knowledge in a Social World'.
Goldman, Alvin, and Erik J. Olsson. 2009. 'Reliabilism and the Value of Knowledge'. In *Epistemic Value*, edited by Adrian Haddock, Alan Millar, and Duncan Pritchard, 19–41. Oxford University Press.
Graham, Peter J. 2010. 'Testimonial Entitlement and the Function of Comprehension'. In *Social Epistemology*, edited by Duncan Pritchard, Alan Millar, and Adrian Haddock, 148–74. Oxford University Press.
———. 2012a. 'Testimony, Trust, and Social Norms'. *Abstracta* 6 (3): 92–116.
———. 2012b. 'Epistemic Entitlement'. *Noûs* 46 (3): 449–82. https://doi.org/10.1111/j.1468-0068.2010.00815.x.

———. 2015. ‘Epistemic Normativity and Social Norms’.
Greco, John. 2009. ‘The Value Problem’. In *Epistemic Value*, edited by Adrian Haddock, Alan Millar, and Duncan Pritchard, 313–22. Oxford: Oxford University Press.
———. 2010. *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge University Press.
———. 2015. ‘Testimonial Knowledge’. *Epistemic Evaluation: Purposeful Epistemology*, 274.
———. 2019. ‘The Transmission of Knowledge and Garbage’. *Synthese*, 1–12.
Green, Christopher R. 2008. ‘Epistemology of Testimony’. *Internet Encyclopedia of Philosophy*, 1–42.
Haddock, Adrian, Alan Millar, and Duncan Pritchard, eds. 2009. *Epistemic Value*. Oxford: Oxford University Press.
Hardin, Russell. 1992. ‘The Street-Level Epistemology of Trust’. *Analyse & Kritik* 14 (2): 152–176.
———. 2002. *Trust and Trustworthiness*. Russell Sage Foundation.
Hawley, Katherine. 2014. ‘Trust, Distrust and Commitment’. *Noûs* 48 (1): 1–20.
———. 2019. *How to Be Trustworthy*. Oxford University Press, USA.
Hieronymi, Pamela. 2008. ‘The Reasons of Trust’. *Australasian Journal of Philosophy* 86 (2): 213–36. https://doi.org/10.1080/00048400801886496.
Hobbes, Thomas. 1970. ‘Leviathan (1651)’. *Glasgow 1974*.
Holton, Richard. 1994. ‘Deciding to Trust, Coming to Believe’. *Australasian Journal of Philosophy* 72 (1): 63–76. https://doi.org/10.1080/00048409412345881.
Horsburgh, H. J. N. 1960. ‘The Ethics of Trust’. *The Philosophical Quarterly (1950-)* 10 (41): 343–54. https://doi.org/10.2307/2216409.
Hume, David. 1739. *Treatise on Human Nature*. Oxford University Press.
Jones, Karen. 1996. ‘Trust as an Affective Attitude’. *Ethics* 107 (1): 4–25.
———. 2004. ‘Trust and Terror’. In *Moral Psychology: Feminist Ethics and Social Theory*, edited by Peggy DesAutels and Margaret Urban Walker, 3–18. Rowman & Littlefield.
———. 2012. ‘Trustworthiness’. *Ethics* 123 (1): 61–85.
Kelp, Christoph, and Simion, Mona. Forthcoming. A Social Epistemology of Assertion (with C. Kelp). *Oxford Handbook of Social Epistemology*, Lackey, J. and McGlynn, A. (eds.). Oxford: Oxford University Press.
——— 2020a. *Giving Knowledge: A Functionalist Account of the Nature and Normativity of Assertion*. MS.
——— 2020b. Trustworthiness as the Disposition to Fulfil One’s Obligations. MS.
——— 2016. The Tertiary Value Problem and the Superiority of Knowledge (with C. Kelp). *American Philosophical Quarterly,* vol. 53/4: 397-411
Keren, Arnon. 2014. ‘Trust and Belief: A Preemptive Reasons Account’. *Synthese* 191 (12): 2593–2615.

Kraut, Robert. 1980. 'Humans as Lie Detectors'. *Journal of Communication* 30 (4): 209–218.
Lackey, Jennifer. 2003. 'A Minimal Expression of Non–Reductionism in the Epistemology of Testimony'. *Noûs* 37 (4): 706–723.
———. 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press.
Lipton, Peter. 1998. 'The Epistemology of Testimony'. *Studies in History and Philosophy of Science Part A* 29 (1): 1–31.
Lyons, Jack. 1997. 'Testimony, Induction and Folk Psychology'. *Australasian Journal of Philosophy* 75 (2): 163–178.
McGlynn, Aidan. 2017. 'Mindreading Knowledge'. In *Knowledge First: Approaches in Epistemology and Mind*, edited by Joseph Adam Carter, Emma C. Gordon, and Benjamin W. Jarvis, 72–94. Oxford: Oxford University Press.
McLeod, Carolyn. 2002. *Self-Trust and Reproductive Autonomy*. MIT Press.
McMyler, Benjamin. 2011. *Testimony, Trust, and Authority*. OUP USA.
Medina, José. 2011. 'The Relevance of Credibility Excess in a Proportional View of Epistemic Injustice: Differential Epistemic Authority and the Social Imaginary'. *Social Epistemology* 25 (1): 15–35.
———. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. Oxford University Press.
Möllering, Guido. 2006. *Trust: Reason, Routine, Reflexivity*. Elsevier.
Nagel, Jennifer. 2013. 'Knowledge as a Mental State'. *Oxford Studies in Epistemology* 4: 275–310.
O'Neill, Onora. 2002. *Autonomy and Trust in Bioethics*. Cambridge University Press.
Origgi, Gloria. 2012. 'Epistemic Injustice and Epistemic Trust'. *Social Epistemology* 26 (2): 221–235.
Owens, David. 2017. 'Trusting a Promise and Other Things'. *The Philosophy of Trust*, 214–29.
Piovarchy, Adam. forthcoming. 'Responsibility for Testimonial Injustice'. *Philosophical Studies*, 1–19. https://doi.org/10.1007/s11098-020-01447-6.
Pohlhaus Jr, Gaile. 2014. 'Discerning the Primary Epistemic Harm in Cases of Testimonial Injustice'. *Social Epistemology* 28 (2): 99–114.
Potter, Nancy Nyquist. 2002. *How Can I Be Trusted? A Virtue Theory of Trustworthiness*. Rowman & Littlefield.
Pritchard, Duncan. 2004. 'The Epistemology of Testimony'. *Philosophical Issues* 14: 326–348.
Rabinowicz, Wlodek, and Toni Ronnow-Rasmussen. 2000. 'II-A Distinction in Value: Intrinsic and For Its Own Sake'. *Proceedings of the Aristotelian Society* 100 (1): 33–51.

Reid, Thomas. 1764. 'An Inquiry into the Mind on the Principles of Common Sense'. In *The Works of Thomas Reid*, edited by W.H. Bart. Maclachlan & Stewart.
Rønnow-Rasmussen, Toni. 2003. 'Subjectivism and Objectivism'. *Rabinowicz W, Rønnow-Rasmussen T (2003) Patterns of Value. Lund: Lund Philosophy Reports* 1.
Simion, Mona. 2020. 'Testimonial Contractarianism: A Knowledge-First Social Epistemology'. *Noûs*.
Simion, Mona, and Christoph Kelp. 2018. 'How to Be an Anti-Reductionist'. *Synthese*, March. https://doi.org/10.1007/s11229-018-1722-y.
Solomon, Robert C., and Fernando Flores. 2003. *Building Trust: In Business, Politics, Relationships, and Life*. Oxford University Press USA.
Sosa, Ernest. 2010a. 'How Competence Matters in Epistemology'. *Philosophical Perspectives* 24 (1): 465–475.
———. 2010b. 'Value Matters in Epistemology'. *The Journal of Philosophy* 107 (4): 167–190.
———. 2015. *Judgment and Agency*. Oxford: Oxford University Press.
Sylvan, Kurt. 2017. 'Veritism Unswamped'. *Mind* 127 (506): 381–435.
Wanderer, Jeremy. 2017. 'Varieties of Testimonial Injustice'. In *The Routledge Handbook of Epistemic Injustice*, 27–40. Routledge.
Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.

Author Information

J. Adam Carter
Email: adam.carter@glasgow.ac.uk
COGITO Epistemology Research Centre, University of Glasgow
United Kingdom

Mona Simion
Email: mona.simion@glasgow.ac.uk
COGITO Epistemology Research Centre, University of Glasgow
United Kingdom

Authors' Note: The authors' research on trust was conducted as part of the Leverhulme-funded 'A Virtue Epistemology of Trust' (#RPG-2019-302) project, which is hosted by the University of Glasgow's COGITO Epistemology Research Centre, and the authors are grateful to the Leverhulme Trust for supporting this research.