Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cphp20

Trust, trustworthiness, and obligation

Mona Simion & Christopher Willard-Kyle

To cite this article: Mona Simion & Christopher Willard-Kyle (2024) Trust, trustworthiness, and obligation, Philosophical Psychology, 37:1, 87-101, DOI: 10.1080/09515089.2023.2223221

To link to this article: https://doi.org/10.1080/09515089.2023.2223221

Published online: 13 Jun 2023.

PHILOSOPHICAL PSYCHOLOGY
2024, VOL. 37, NO. 1, 87–101
https://doi.org/10.1080/09515089.2023.2223221

Routledge
Taylor & Francis Group

Trust, trustworthiness, and obligation

Mona Simion and Christopher Willard-Kyle

Cogito Epistemology Research Centre, University of Glasgow, Glasgow, UK

Abstract

Where does entitlement to trust come from? When we trust someone to φ, do we need to have reason to trust them to φ or do we start out entitled to trust them to φ by default? Reductivists think that entitlement to trust always "reduces to" or is explained by the reasons that agents have to trust others. In contrast, anti-reductivists think that, in a broad range of circumstances, we just have entitlement to trust. even if we don't have positive reasons to do so. In this paper, we argue for a version of anti-reductivism. Roughly, we argue that we have default entitlement to trust someone to φ so long as there is an operative norm that requires S to φ. At least in such circumstances (and absent defeaters), we don't need any positive reasons to trust S to φ.

ARTICLE HISTORY
Received 27 September 2022
Accepted 31 May 2023

Keywords
trust; trustworthiness; obligation; norm

I. Introduction

When should we trust people to φ? Presumably, we should trust people to φ when they are *trustworthy* with respect to φ-ing. Indeed, paradigmatically good instances of trusting involve the trust of the truster *matching* the trustworthiness of the trustee (Carter Forthcoming, O'Neill, 2018). Suppose I entrust George with watching my dog when I'm on holiday. But George is *not* trustworthy with respect to watching my dog, and in fact only manages to do so because his parents go to great lengths to ensure that he does. Then even though George has done what I have entrusted him to do, my trusting him has not been fully successful. Part of the reason for this is that the strength of my trust in George to watch my dog is not matched by George's actual level of trustworthiness with respect to watching my dog.[1]

The ideal situation, then, would be that we trust people to φ just in those cases that they are trustworthy with respect to φ-ing. Of course, the ideal is often hard to achieve. When actually making decisions about whom to trust,

CONTACT Mona Simion mona.simion@glasgow.ac.uk Cogito Epistemology Research Centre, University of Glasgow, 69 Oakfield Avenue, Glasgow G12 8LP, UK

Invited for Special Issue of Philosophical Psychology on "Trustworthiness: Individual and Institutional Dimensions", Eds. Elisabetta Lalumera and Filippo Ferrari

we have to make do with the information we have. We might not be entitled to trust someone even if they are in fact trustworthy, *e.g.*, because we have misleading evidence that they are dishonest. We might also be entitled to trust someone when they really aren't trustworthy, as is the victim of a well-executed con.

Here's a natural question that arises given this setup. Where does entitlement to trust come from? When we trust someone to φ, do we need to have *reason* to trust them to φ or do we start out entitled to trust them to φ by default? Do we need reasons to trust people or not?

Let us introduce some terminology, borrowing from the epistemology of testimony. We can frame this question as a debate between *reductivists* about entitlement to trust and *anti-reductivists*. Reductivists think that entitlement to trust always "reduces to" or is explained by the reasons that agents have to trust others. In contrast, anti-reductivists do think that our entitlement to trust is not generally explained by reasons to trust. In a broad range of circumstances, we just have entitlement to trust even if we don't have positive reasons to do so. Do we get entitlement from *reasons* (reductivism) or by *default* (anti-reductivism).

This turns out to be a hard question. The reason it's hard is that, at first glance, both answers seem obviously wrong. The anti-reductivist seems unable to explain the thought that trust is *earned*. Especially when we entrust someone with an important task, it seems foolhardy not to do so on the basis of evidence. Imagine dropping off your children at a stranger's house without having first done some serious research! Trusting, and thereby risking, without positive reasons to trust the trustee seems reckless. Anti-reductivism also threatens to make us gullible. We aren't entitled to trust QAnon conspiracy theorists just because they say they've done their research. Anti-reductivists are suckers. Or so the worry goes.

On the other hand, reductivism looks no better: reductivism seems too strong. Indeed, certain paradigmatic instances of (appropriate) trust apparently happen without any positive reasons to trust whatsoever. Children trust their parents to take care of them, but young children aren't cognitively sophisticated enough to appreciate reasons for thinking that their parents are trustworthy. I trust the driver at the intersection not to blow through the red light, but I've never even met them, and I have no reason for thinking that they, in particular, are trustworthy with respect to following the rules of the road. Stubbornly refusing to trust in such circumstances would be immobilizing. Reductivists are too hesitant to trust: reductivists, it seems, are skeptics.[2]

We face an apparent dilemma. Reductivists are skeptics; anti-reductivists are suckers. We don't want to be either.

In the rest of the paper, we explore this apparent tension and chart a new path forward. We argue for a version of anti-reductivism. Roughly, we argue that we have default entitlement to trust someone to φ so long as there is an operative norm that requires S to φ. At least in such circumstances (and absent defeaters), we don't need any positive reasons to trust S to φ.

II. Trustworthiness to φ and obligation

Before diving into our argument, let's establish some context and make some distinctions.

Anti-reductivists can be *modest* or *extreme*. Extreme anti-reductivists think one *always* has default entitlement to trust S to φ (absent defeaters). Modest anti-reductivists think that one at least *sometimes* has default entitlement to trust S to φ (absent defeaters). We are modest anti-reductivists. Our anti-reductivism is modest because we don't claim that one has a default entitlement to trust no matter what: One has default entitlement to trust someone to φ only if a norm to φ is operative.

This isn't by accident. We think there's a substantive relationship between norms and trust that manifests both in the conditions for trustworthiness and in the conditions for entitlement to trust. That is to say, entitlement to trust and trustworthiness are both explained, in some way, by norm-compliance.

One point of clarification: when we talk about norms in this paper, we employ the broad, philosophical, rather than the more narrow, sociological understanding thereof: that is, on the account henceforth employed, norms are abstract objects that need not be embedded in social practices.[3]

In explaining entitlement to trust and trustworthiness in terms of norm-compliance, we follow mainstream work in the literature on trustworthiness in thinking that being trustworthy has to do with one having a disposition to do what one is supposed to do. There are many incarnations of this general thought defended in the literature: For example, according to a family of views defended by Annette Baier (1986), Karen Jones (2012), and Zac Cogley (2012), trustworthiness is to be identified with a disposition to fulfil commitments, in conditions under which one has those commitments, and in virtue of goodwill toward the trustor. For Diego Gambetta (1988), the trustworthy person needn't be disposed to fulfill the commitments they have out of good will; they simply must be disposed to fulfil their commitments, whatever they are, "willingly". More minimalistically, (Kelp & Simion, 2023) identifies trustworthiness with the disposition to fulfill one's obligations simpliciter, and not necessarily through any distinctive motivation or accompanying attitude. More weakly, for Katherine Hawley (2019), the relevant disposition referred to is best framed negatively – viz., as a disposition to avoid unfulfilled commitments. By contrast, more strongly, according to Nancy Potter (2002), the relevant disposition lining up with trustworthiness should be understood as a full-fledged moral virtue – one that consists in being disposed to respond to trust in appropriate ways.[4]

For our purposes here, we will follow the account defended by one of us in previous work (Simion & Kelp 2023a). We think the account has several advantages over the competition – including extensional adequacy, as well as generalizability e.g., to understanding trustworthy institutions, or trustworthy AI agents (Simion & Kelp, 2023). That being said, not much will hinge on this, any account that vindicates this more general thought – i.e., that trustworthiness concerns being disposed to do what one is supposed to do – will do for the purposes of this paper. Importantly, though, we will not offer a full defense of the account here – it falls outside the scope and requirements of the argument made in this paper.

The view proposes to make sense of what it is for someone to be trustworthy with regard to a particular action φ by starting with an account of maximal trustworthiness to *phi*. More specifically, it starts with the following intuitively highly plausible idea: to have the property of trustworthiness to φ to its fullest (henceforth also maximal trustworthiness to φ) is to as strongly disposed to fulfil one's obligations to φ as possible:

Maximal trustworthiness to φ

One is maximally trustworthy with regard to φ-ing if and only if one has a maximally strong disposition to fulfil one's obligations to φ.

For instance, according to Maximal Trustworthiness to φ, to be maximally trustworthy when it comes to doing the dishes is to have a maximally strong disposition to wash the dishes when under an obligation to wash the dishes.

Dispositions may vary in degree of strength. The higher the probability of manifestation given presence of the trigger (i.e., relevant obligation) in suitable conditions,[5] the stronger the disposition.[6]

Maximal Trustworthiness to φ states necessary and sufficient conditions for maximal trustworthiness to φ. At the same time, we human beings are finite and so we are rarely if ever in the ballpark for maximal trustworthiness to φ. Nevertheless, we frequently attribute trustworthiness to φ to each other. How can we make sense of our practice of attributing trustworthiness to φ? To answer this question, (Simion & Kelp 2023a) first offers the following account of degrees of trustworthiness to φ:

Degrees of trustworthiness to φ

The degree of trustworthiness to φ of S is a function of the distance from maximal trustworthiness to φ: the closer one approximates maximal trustworthiness to φ, the higher one's degree of trustworthiness to φ.

Suppose that while George is generally disposed to live up to his obligation to wash the dishes, he may fail to do so when the Eurovision finals are on or when he is about to finish the book he is reading. Ann is also generally disposed to live up to her obligation to wash the dishes. She may fail to do so when the Eurovision finals are on, but she will not let an almost finished book get in the way. Degrees of Trustworthiness to *Phi* predicts that Ann is more trustworthy when it comes to doing the dishes than George is – which is also the right result, intuitively.[7]

Next, we combine this account of degrees of trustworthiness to φ with a contextualist semantics for outright attributions of trustworthiness to φ According to this account of outright attributions of trustworthiness to φ, context determines a threshold on degrees of trustworthiness to φ such that one is trustworthy to φ just in case one surpasses the threshold in question. Or, to be more precise,

> **Attributions of outright trustworthiness to φ**
>
> "*S* is trustworthy to φ" is true in context *c* if and only if *S* approximates maximal trustworthiness to φ closely enough to surpass a threshold on degrees of trustworthiness determined by *c*.

On this view, then, when, at a particular context, we say that Ann is trustworthy when it comes to washing the dishes but George isn't, what is happening is that Ann approximates a maximally strong disposition to do so, conditional on having the corresponding obligation, to a contextually sufficiently high degree, whereas George doesn't. Just how high the threshold is will be determined by and may vary with context. To see that this is plausible, compare a case in which Ann and George are professional dishwashers at a local restaurant with a case in which they are Mary's teenage children. It is intuitively plausible that the threshold for what it takes to count as trustworthy when it comes to washing the dishes is higher in the first case than in the second.

Notice that this definition is only about trustworthiness to φ and not trustworthiness *simpliciter* (Jones, 1996).[8] Second, the obligations that are invoked do not *have* to be ethical obligations. So, a professional thief could be maximally trustworthy with regard to fulfilling contracts to procure illicit goods even if (for the very same reason) they are not trustworthy *simpliciter*. Third, notice that on this definition, one cannot be (non-trivially) trustworthy with respect to φ-ing unless one has an obligation to φ. Obligations occasion the possibility of (non-trivial) trustworthiness.[9]

Norms are often obligation-generating. If one is maximally disposed to satisfy a certain obligation-generating norm, then one is also maximally disposed to fulfill one's obligation to comply with the relevant norm. In that way, facts about trustworthiness are partially explained by facts about norms. Obligation-generating norms create conditions of obligation, and dispositions to satisfy those obligations constitute degrees of trustworthiness.

Our central idea is that, just as norms play a starring role in shaping the contours of trustworthiness, so norms play a starring role in shaping the contours of entitlement to trust. Those cases in which one has *default* entitlement to trust are those in which an obligation-generating norm is operative. What is default entitlement to trust someone? It is entitlement to trust in the absence of any had evidence for or against trusting, as well as in the absence of any had non-epistemic reason for or against trusting.

Crucially: In line with the definitions above, we emphasize that we are defending an anti-reductivism about trusting someone to φ rather than trusting someone *simpliciter*. That is, our thesis is about trust understood as a three-place relation between a truster, trustee, and some action φ-ing rather than trust as a two-place relation between a truster and a trustee. Arguably, the conditions for two-place trust are more demanding than three-place trust; importantly, we're *not* committed to the view that (*e.g.*) Ann doesn't need reasons to trust George (*simpliciter*). Our argument is silent on the subject.

Here's another distinction that matters for our argument. We're arguing that, in certain cases, one can have *default* entitlement to trust someone to φ. Default entitlement is defeasible. So we're not saying that one can always trust others to comply with the norms. One is flat-out entitled to do so when one's default entitlement is not ultimately defeated.[10]

Defeat comes in different flavors. Importantly, it's not just the case that one's entitlements are defeated by reasons one *has*. They can be defeated by reasons one *ought* to have, the things one *should have* known (cf. Goldberg, 2017). These are *normative* defeaters. Defeaters defeat default entitlement, so our thesis is just that one (still) has entitlement to trust someone to φ when there is an operative norm to φ *and no defeaters are present*.

The kind of entitlement that is at issue in our argument is epistemic and not, importantly, moral entitlement. Recall that a thief can be trustworthy at thieving even though this kind of trustworthiness isn't morally valuable. What goes for trustworthiness goes for entitlement to trust. Entitlement to trust can also be morally disvaluable. For example, a father might trust that his daughter will abide by certain gender norms. This trust can be rationally entitled for him (*i.e.*, if it's true and supported by his evidence that his daughter does indeed abide by the relevant norms) even if such trust is morally irksome.

Let's recap. We are arguing for a form of anti-reductivism for epistemic entitlement to trust. The anti-reductivism is *modest* in that it only says that we are *sometimes* entitled to trust without reason. The entitlement is *default* in that it is defeasible. And the relevant kind of trust is a three-place relation between a truster, trustee, and φ-ing. Anti-reductivism about entitlement to trust is our general thesis, but here's the specific version of it we like:

> NORM-BASED, MODEST ANTI-REDUCTIVISM (NOMAR): If in a given context, there is an operative norm such that S_1 is obligated to φ, then S_2 has default entitlement to trust S_1 to φ.[11]

We argue for this thesis below.

III. Compliance contractarianism

Our argument relies on a certain kind of contractarianism. On strong contractarian views, some story or other about the social contract explains the very existence of (*e.g.*) moral or political norms. We're not concerned with strong contractarianisms of this kind. What we *are* interested in is the more modest contractarian thesis that the social contract in fact motivates people to comply with operative norms even when defecting from those norms would serve their own self-interest. Norms work – people don't (generally) cut queues at storefronts even when it'd be in their interest to do so. Call this thesis *Compliance Contractarianism* (Simion, 2021b; Kelp & Simion 2021).

Contractarianism finds its historical origins in Thomas Hobbes (1651, 1651). Here's the basic picture. We begin by positing a (perhaps merely hypothetical) "state of nature". In this state of nature, people are unconstrained to pursue their own self-interest. This apparent freedom leads to bad results, however, because the self-interest of different selves conflict. And without any constraints on behavior, people in the state of nature are free to harm each other in pursuit of their own advantage. This is predictably bad. And so, life in the state of nature proves to be – in Hobbes' (1651, XIII) stark but memorable line – "nasty, brutish, and short".

Because life in the state of nature – life in a world without norms that constrain us – is so inhospitable, it's in our rational interest to trade a degree of our freedom for the benefits that come from a norm-governed social order. But there's a social coordination problem. It is only in our interest to trade our freedom for the benefits of a norm-governed social order if the whole group makes the same trade. To leave this state of nature, what we need is a group policy – a 'social contract'—with those in our community to abide by certain norms. These norms constrain our ability to seek our own self-interest, but (at least when the norms are good ones) we are each better off when we all follow the contractual norms.[12]

That's the basic story. One important set of questions concerns getting more detail about *how* these norms come into place: how do we collectively decide which and how many freedoms to give up in a fair way? But our interest in contractarianism is further down the line. *Once* the norms have been set, do they generally work at constraining group behavior?

There are good reasons to think the answer is yes. The first reason is boringly straightforward. We just observe that lots of norms are, in fact, regularly followed, even when it's costly for agents to comply with hem. Drivers generally don't run red lights even though running lights that have *just* turned red would be more efficient. Parents generally don't send their children to school when they know them to be contagious despite the resulting scheduling hassle. People generally don't cut the queue at the grocery checkout line even though it'd be faster to.

Of course, there are exceptions. These norms are not inviolate. But people follow these norms reliably enough that they reasonably ground an entitlement to trust that they will be followed. Indeed, when social norms aren't followed regularly enough, they tend to evaporate. The erstwhile norm to address fellow adults with "Mr." or "Ms." before their surname is no longer a norm (or at least well on the way out) in part because it simply isn't practiced reliably enough to maintain its status as a default expectation. In this way, the continued existence of a social norm can itself be some (defeasible) evidence that it is reliably enough followed.

A second reason to think that norms tend to stop people from defecting is the possibility of punishment. Some punishments are, of course, administered by the state through the legal system, but we're using "punishment" in a much broader way: honking horns, cold shoulders, and pointed glares all qualify. Punishments enforce social norms by making it costly to defect from the social contract. Once a norm is adequately enforced, it's in the rational self-interest of an agent to abide by the norm.

But that's not to say that we think norms always or only operate because of the possibility of punishment. A third reason that norms can work is that they are habit-forming. We doubt that the last time you stood in a queue, you deliberated about whether the time you would save by cutting the queue would offset the irritated glares or uncomfortable conversations you would encounter by so cutting. Going to the back of the queue is just the thing to do. You got in the back of the queue as part of a default policy.

But perhaps most importantly, there's empirical evidence from psychology that we do in fact enforce social norms, even when doing so conflicts with our self-interest. Consider ultimatum games that take the following form:

> ULTIMATUM: This game has two players, the *proposer* and the *responder*. The game begins with the introduction of monetary stakes. An amount of money is named (*e.g.*, $100). Then the proposer makes the following move: they propose a way of dividing the money between themselves and the responder. For instance, they could propose that they get $100 and the responder gets $0. Or they could propose that they get $20

> and the responder gets the remaining $80, and so on. After the proposer's move, it's the responder's turn. They can make one of two moves: accept the proposal or reject it. If they accept the proposal, both the proposer and the responder walk away with the amount of money stipulated in the proposal. If the responder rejects the proposal, the proposer and the responder both walk away with nothing.

ULTIMATUM is not an iterated game. From a purely game-theoretic perspective, it seems like the responder should accept any non-zero sum of money. After all, if they accept they will get something whereas if they reject they will get nothing. Recognizing that it's rational for responder to accept any non-zero offer, it seems that proposer should (in order to maximize their own earning) make a proposal that gives the smallest possible increment to the responder, *e.g.*: $99 for the proposer, $1 for the responder.

But this is very rarely how things play out. Instead, proposers tend to offer something closer to an even split – something around 60/40. And when proposers make lopsided proposals (as traditional game theory seems to suggest they should), responders tend to reject rather than accept the proposals. Neither player plays like they're supposed to.

What's going on in these cases? Here's what we think is happening. When responders reject lowball offers, they're enforcing a norm to treat others fairly. The lowball offer is not perceived as fair – after all, the cooperation of both participants is required to secure any money. Responders view enforcing these norms as worth more than the pittance they'd get from accepting the lowball offer.

Proposers implicitly recognize this, and so they *don't* reason (as in the traditional, game-theoretic story above) that the responder will accept any non-zero offer. The offer must be high enough to outweigh the cost (from the responder's perspective) of failing to enforce the norm.

Importantly, the norms make a difference to the expected outcomes of the players' decisions. The proposer must rationally take into account whether their proposal is norm-abiding (enough) when deliberating about how responder will react and (ultimately) what to do. What the case seems to show is that it's rational to expect people to abide by (and enforce) the norms even when doing so is out of line with their immediate self-interest.

Predictably, when the stakes are raised – when the cost for the responder of rejecting, say, 10% of the pot goes up because the pot is bigger – responders are less eager to enforce the norms:

> [A]mong respondents we find a considerable effect of stakes: while at low stakes we observe rejections in the range of the extant literature, in the highest stakes condition we observe only a single rejection out of 24 responders. (Andersen et al., 2011)

This isn't surprising on our account. Default considerations can be defeated or overridden. People tend to enforce norms and assign real weight to enforcing them in the utility profile for their decision-making. Make the cost for enforcing norms high enough, and people will eventually stop enforcing them. But when people aren't paid off to defect, they tend assign real weight to enforcing operative social norms within their utility profile. Because we are constantly surrounded by motivated norm-enforcers, the default rational position from within a community governed by a social contract is to comply with the norms.[13]

Still, one might wonder: even if norms generally work, is it correct that people are motivated to comply with them reliably? Is it plausible that this will hold for all people, or most – which seems to be what the account requires? Three things about this: first, we, to a large extent, assume Compliance Contractarianism holds for the purposes of this paper: it is not our ambition to offer a novel defense thereof, plenty of work has been done in the literature; second, we have given empirical support for the claim that vast norm compliance is in effect, absent reason against, and (3) finally, note that the account works even if we restrict the motivational claim to rational agents: in the absence of any motivation to break the norm, the existence of the norm itself renders norm compliance the dominant option. On the further assumption that one is entitled to assume agents with rational capacities will act rationally – absent defeat –, NoMar holds.

Of course, sometimes we have evidence that particular norms are not followed. There's a norm, etched in law, that pedestrians should only cross streets at designated crossings. But these norms are routinely violated on small or quiet streets. Our view does not have the implication that we have entitlement to trust pedestrians not to step into the street in the relevant circumstances. That's because our knowledge that people don't reliably follow the norm defeats our default entitlement. We don't need reasons to believe a particular passerby has a reason to break the norm: a fairly minimal account of defeaters as facts that decrease evidential probability will deliver the result that reason to believe that norms demanding *phi*-ing are reliably enough broken will defeat entitlement to trust any particular subject to phi.

What do "routinely" and "reliably enough" stand for? This is a version of the classic threshold problem for infallibilism: insofar as one thinks reliability is enough for justification/infallibility is not necessary, the question as to where to set the relevant threshold becomes relevant. We don't aim to try to answer this question here – indeed, we are skeptical about the availability of a precise answer for any normative domain, and even more so about entitlement to trust, which transcends normative domains – it can be epistemic, practical, moral etc, depending on context and what is at stake. Context, stakes, and normative domains will weigh heavily here: I have a default entitlement to trust pedestrians in Glasgow to only step into the street at designated crossings insofar as this is an active norm in Glasgow and Bucharest. If, however, I find out that people in Glasgow do cross the street in breach of this norm all the time (because, say, the norm is not legally enforced), my entitlement is defeated. My entitlement can also be normatively defeated: if I should have known that pedestrians in Glasgow do this routinely (say, because I lived in Glasgow for 6 years now), knowledge that I should have had also defeats my entitlement to trust pedestrians in this respect. How "routinely" do pedestrians need to do this for my norm-generated entitlement to be defeated? It depends on the type of entitlement, and the issue at stake. If we're talking about epistemic entitlement to trust, the number of violations required for full entitlement defeat will be predicted by the correct reliabilist view about the threshold for epistemic justification. If it is moral or practical entitlement that we are talking about, and given that lives are at stake, the threshold will likely be higher: the expected disutility of seriously injuring a passerby is high, so even a small probability of norm violation affects entitlement.

We've now given some reason to believe that compliance contractarianism is plausible. In general, norms work. When there is an operative norm in place, people are in fact motivated to comply with the operative norm reliably. Now, recall our central thesis:

> NORM-BASED, MODEST ANTI-REDUCTIVISM (NOMAR): If in a given context, there is an operative norm such that S_1 is obligated to φ, then S_2 has default entitlement to trust S_1 to φ.

NOMAR says that, when norms are present, entitlement to trust is cheap. Compliance contractarianism explains why. When norms to φ are present, the social contract kicks in, making it likely enough across a wide range of cases that the requisite φ-ing will happen. Operative norms can make it reasonable for agents to behave in ways that would otherwise (in the absence of such norms) violate their self-interest. Accordingly, self-interested agents have reason to comply with norms rather than to defect. Indeed, they tend to do so as a matter of default.

Since it's likely enough that agents will comply with the norms, trusting agents to φ when there is an operative norm to φ is reliable enough. Of course, the pressures to conform to the social contract can be overridden, and likewise the entitlement to trust others to fulfil the social contract can be defeated. But the default pressure to abide by the contract creates default entitlement to trust others to do so.

In passing, we add that NOMAR also adds to the explanation of why it can be so frustrating when someone breaks the social contract in mild but flagrant ways. Cutting the queue at the grocery store is pretty harmless. Having to wait an extra minute to get your groceries is pretty far down the list of bad things that can happen to you. But queue-cutting and other mild but flagrant violations of the social contract can produce outrage that is out of proportion with the harm caused. (If you don't believe us, just try it!) This can seem surprising if we only focus on the harm caused by such infractions.

But with NOMAR in mind, it isn't surprising at all. Such infractions are flagrant violations of trust.

We've been arguing that compliance contractarianism is evidence for moderate anti-reductivism about entitlement to trust. A reductivist might try to respond this way:

> You haven't offered an explanation of anti-reductivism at all! After all, you've given us a *reason*—namely, compliance contractarianism—for thinking that it's a good idea to trust people to φ when we also know there is a norm to φ. And giving reasons to trust is the trade of the reductivist. The explanation for why you can trust generally trust people to follow the social norms isn't that we have default entitlement in such cases, it's that we have a reason—compliance contractarianism—that entitles us to trust people to φ.

There is, however, an important distinction between there being reason to trust someone, and one having reason to trust someone (Simion 2023).[14] What we have given – and indeed, what all anti-reductivist champions have offered in the literature – are reasons to believe anti-reductivism is true, and thereby reasons to believe trustors have a default entitlement to trust. Compatibly, on our account, and in contrast to reductivist views, the trustors do not need to *have* these reasons themselves. We think requiring trustors to understand (even implicitly) complicated philosophical anti-reductivist accounts is far too high brow. People don't need to take a course on social contract theory to have entitlement to trust – nor do they even have to have a rough implicit understanding thereof. Entitlement to trust in social contexts is not restricted to economists and political scientists, and certainly not to philosophers. Furthermore, philosophers who oppose contractarianism – i.e., the paradigmatic agents who will not have the reasons we're offering (no matter what account of "having" we employ) in virtue of nut buying into our theory – can have entitlement to trust. If compliance contractarianism plays a role in entitling us to trust others to comply with the norms, it is not, in the first instance, by being a reason on the basis of which we conclude that others will abide by the norms. Rather, it directly entitles us to trust by making our trust reliable enough.

IV. Conclusions

We began with a dilemma. Reductivistm and anti-reductivism both seemed like nonstarters. Reductivism made it hard to explain how entitlement to trust could be appropriately unsophisticated: young children and adults, too, often automatically trust others to follow through on their obligations in a wide range of social situations. Anti-reductivism seemed unable to avoid the charge of gullibility.

Modest anti-reductivism is a promising path forward. We are entitled to trust others to φ in certain cases – those in which there is a norm requiring others to φ. Compliance contractarianism ensures that operative norms are, in general, reliably enough followed to license default trust to φ without succumbing to gullibility.

Modesty has its costs. An important challenge for the modest anti-reductivist is to distinguish those cases in which one has default entitlement from those in which one does not. But our way of drawing this line is demonstrably not *ad hoc*. Indeed, it flows out of the account of trustworthiness we started with, and, arguably, versions of this view will follow on any account that takes trustworthiness to have to do with a disposition to meet one's obligations. Trustworthiness to φ just is the disposition to satisfy the norms to which one is bound. It's no surprise, then, that norms should also figure in the explanation of when we are entitled to trust others to φ. Trustworthiness is a normative, through and through.

Notes

[1] For elaboration, see Carter (Forthcoming).

[2] See e.g (Kelp & Simion, 2017) for an account of the distinctive value of knowledge in terms of easy availability.

[3] Thanks to an anonymous referee for pressing us to clarify this.

[4] For an extensive overview, see (McLeod, 2015) and (Carter & Simion, 2020).

[5] For what we take to be a compelling case that dispositions are relative to suitable conditions, see (Mumford, 1998) and (Sosa, 2015).

[6] For more on probabilistic approaches to dispositions see (Healey, 1991) and (Suarez, 2007).

[7] Why not an account of trustworthiness in terms of observed dispositions? The main reason against such an account is that the manifestation of dispositions depends on environmental conditions. I might make it to lunch as promised only 9 times out of 10, while Sally makes it all 10 times. However, if this happened because my neighborhood was placed in lockdown suddenly, I am not less trustworthy than Sally. Our account predicts, correctly, that Sally and I are equally trustworthy.

[8] For an account of how to build an account of Maximial Trustworthiness *simplliciter* from Maximal Trustworthiness to φ, see Kelp and Simion (2023).

[9] One great advantage of this view of trustworthiness is that, in contrast with the vast majority of its competition, it is not anthropocentric: any entity governed by obligation-generating norms can be trustworthy. In a different paper, one of us spells this out in detail for artifacts, and in particular for AIs: artifacts have design functions, and often even etiological functions. – i.e., functions having to do with what they were (socially) selected for. Design functions and etiological functions notably generate norms of proper functioning: the artifact in question will be properly functioning or malfunctioning depending on whether it works in the way in which, in normal conditions, it reliably enough fulfills

its function. In this sense, the artifact in question ought to work in a reliable-function-generating manner. See (Simion and Kelp, 2023) for a full defense of the view.

[10] For those who traffic in the relevant ideology, the "ultimately" qualifier in "ultimately defeated" is intended to account for the possibility of defeater-defeaters. A default entitlement can have a defeater. But if that defeater is itself defeated by a defeater-defeater, then the default entitlement is not *ultimately* defeated.

[11] We don't defend the necessity direction of this claim because we are not convinced it holds: it seems to us as though one can also have entitlement to trust in the absence of trustworthiness-making features – for instance, in the presence of laws of nature that ensure the trustee will do what they are entrusted to do. There is, of course, an interesting question here as to whether this is genuine entitlement to trust, or rather mere entitlement to reply upon. We plan to investigate this in further work.

[12] See Faulkner (2007) for discussion of rational trust and self-interest.

[13] This argument expands on an argument for norm compliance that appears in Simion (2021b: 908–09). See also (Kelp and Simion, 2021a).

[14] Many thanks to two anonymous referees for pressing us on this.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the H2020 European Research Council [948356]; Leverhulme Trust [A Virtue Epistemology of Trust]

ORCID

Christopher Willard-Kyle http://orcid.org/0000-0002-3783-1073

References

Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes matter in ultimatum games. *The American Economic Review*, *101*(7), 3427–3439. https://doi.org/10.1257/aer.101.7.3427

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

Carter, J. A., & Simion, M. (2020). The ethics and epistemology of trust. *Internet Encyclopaedia of Philosophy*.

Cogley, Z. (2012). Trust and the trickster problem. *Analytic Philosophy*, *53*(1), 30–47. https://doi.org/10.1111/j.2153-960X.2012.00546.x

Faulkner, P. (2007). A genealogy of trust. *Episteme*, *4*(3), 305–321. https://doi.org/10.3366/E174236000700010X

Gambetta, D. (1988). *Trust: Making and breaking cooperative relations*. Wiley-Basil Blackwell.

Goldberg, S. C. (2017). Should have known. *Synthese*, *194*(8), 2863–2894. https://doi.org/10.1007/s11229-015-0662-z

Hawley, K. (2019). *How to be trustworthy*. Oxford University Press.

Healey, R. (1991). *The philosophy of quantum mechanics: An interactive interpretation*. Cambridge University Press.

Hobbes, T. (1651). *Leviathan*. C. B. Macpherson, (Ed.). Penguin Books. (1985)

Jones, K. (1996). Trust as an affective attitude. *Ethics*, *107*(1), 4–25. https://doi.org/10.1086/233694

Jones, K. (2012). Trustworthiness. *Ethics*, *123*(1), 61–85. https://doi.org/10.1086/667838

Kelp, C., & Simion, M. (2017). Commodious knowledge. *Synthese*, *194*(5), 1487–1502. https://doi.org/10.1007/s11229-015-0938-3

Kelp, C., & Simion, M. (2021). *Sharing knowledge: A functionalist account of assertion*. Cambridge University Press.

Kelp, C., & Simion, M. (2023). What is trustworthiness? *Nous*. Online First. https://doi.org/10.1111/nous.12448

McLeod, C. (2015). Trust. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/fall2015/entries/trust/

Mumford, S. (1998). *Dispositions*. Oxford University Press.

O'Neill, O. (2018). Linking trust to trustworthiness. *International Journal of Philosophical Studies*, *26*(2), 293–300. https://doi.org/10.1080/09672559.2018.1454637

Potter, N. (2002). *How can I be trusted? A virtue theory of trustworthiness*. Rowman & Littlefield.

Simion, M. (2021a). *Shifty speech and independent thought*. Oxford University Press. https://doi.org/10.1093/oso/9780192895288.001.0001

Simion, M. (2021b). Testimonial contractarianism: A knowledge-first social epistemology. *Nous*, *55*(4), 891–916. https://doi.org/10.1111/nous.12337

Simion, M. (2023). Resistance to evidence and the duty to believe. *Philosophy and Phenomenological Research*. Online First. https://doi.org/10.1111/phpr.12964

Simion, M., & Kelp, C. (2023). Trustworthy artificial intelligence. *Asian Journal of Philosophy*, *2*(1), Special Issue, Ed. N. Pedersen. https://doi.org/10.1007/s44204-023-00063-5

Sosa, E. (2015). *Judgment and agency*. Oxford University Press.

Suarez, M. (2007). Quantum Propensities. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *38*(2), 418–438. https://doi.org/10.1016/j.shpsb.2006.12.003