Philosophical Quarterly Special issue in honour of Katherine Hawley ed. Jessica Brown 02/2023

Abstract: I consider Katherine Hawley's *commitment account* of interpersonal trustworthiness alongside her sceptical challenge regarding the value of philosophically modelling institutional trustworthiness as distinct from reliability. I argue, *pace* Hawley's challenge, that there would be significant diagnostic and explanatory loss if we were to content ourselves with ideas of institutional (un)reliability alone; and I offer an illustrative case where institutional unreliability is only the half of it, indicating that when it comes to certain kinds of institutional dysfunction, we do need philosophical models of institutional 'distrustworthiness' if we are to achieve a proper diagnosis.

Keywords: institutional trustworthiness, distrust, false confession, U.S. police training, Reid Technique, testimonial injustice, agential testimonial injustice

Diagnosing Institutionalized 'Distrustworthiness'

> *Just as we should distinguish trustworthiness from mere reliability, we should distinguish untrustworthiness from mere unreliability… 'Distrustworthy' would be more literal, but I will resist the neologism* -- Katherine Hawley[1]

One of the foci of Katherine Hawley's work is trust, or as she generally preferred, trustworthiness. Her crystal clear and thoroughly engaging book on it in the *Very Short Introduction* series from Oxford University Press introduced a wide readership to key philosophical questions through attention to areas of our social world—personal, professional, political—that depend on trustworthiness. She also wrote a number of research papers on the subject. Her elegant account is what she called a *commitment account*, so that trusting someone to do something should be understood in terms of *believing they will fulfil their commitment to do it*. Her conception of commitment is objective in two respects: one might have a commitment to do something without believing oneself to be so committed; and/or without being motivated to fulfil it.

[1] Hawley 2014 p. 3.

Hawley was consistently attentive to the importance of institutions, often discussing trust in relation to them. She notes, for instance, that we naturally speak of trusting (or distrusting) certain of our institutions: 'We trust institutions—the BBC, Oxfam, the local school, the supermarket and its Fairtrade certification, the Royal Mail, the currency system' (Hawley 2012, p. 28). Although she was quite ready to talk in ordinary ways about trusting institutions, and the social importance of having institutions that are indeed trustworthy, still she was sceptical that this way of talking really added anything that is not already present in the idea of institutions being reliable or unreliable. Rather than argue for a philosophical account of interpersonal trustworthiness that could also do service in relation to institutions, then, she was drawn instead to advance a somewhat sceptical view that, when it came to institutions and groups, little or no value would be lost if we had at our disposal only the diagnostic and critical tools already afforded by the concepts of reliability and unreliability. Accordingly, in Hawley's picture of trust there is no deliberate design for group trustworthiness, and relatedly she considers it no constraint on accounts of collective agents that they should be represented as capable of trustworthiness or its opposite. How could that really matter, after all? Why would we mind whether a reliable institution lacked whatever it might be that would enable the citizenry to depend upon it in the psychologically intricate manner that philosophers have come to characterize as distinctively that of trust? In the introduction to her 2017 paper that focuses exclusively on this issue, Hawley sums up her argumentative strategy as follows:

> I examine the costs of abandoning the trust–reliance distinction in collective contexts. Vindicating this distinction has been regarded as an essential criterion of success for accounts of trust in individuals. But, I will argue, we can explain and justify much of our practices around groups without using this distinction; the costs of abandoning it are low, as compared to the individual case (Hawley 2017, p. 233).

It is this last claim I would like to challenge; and it is the bitter-sweet burden of this paper to present some arguments that in my imagination I am running by Katherine, keen to know her response and hoping to give her pause for thought—always aware that the more likely outcome would have been that she lobbed back some counterarguments that gave me pause instead. Still, I believe she would have acknowledged the powerful phenomenological impetus in favour of pursuing the thought, for there are some (and only some) cases of institutional unreliability that are apt to inspire that special kind of disappointment that is the

hallmark of disappointed trust as opposed to mere reliance—a sense of betrayal. When the London Metropolitan Police is revealed to have long been recruiting high numbers of officers with records of violence and hate that are supposed to prevent them getting through even the most basic background checks, we do not merely roll our eyes in disappointed reliance, but stare in stunned betrayal at the systemic and ethically calamitous dysfunction of one of our most powerful social institutions, funded by the tax payer to serve the public.[2] Spurred by Hawley's challenge regarding the value of theorising institutional trustworthiness as distinct from reliability, I will defend the intellectual value of modelling specifically institutional trustworthiness and 'distrustworthiness' (to adopt her suppressed neologism), on the grounds that we need to be able to diagnose institutional *distrustworthiness* where we confront it. A general observation conducive to this trust and distrust oriented conception is that institutions tend to be powerful, so that we often stand in a relation of significant *dependence* regarding institutional bodies—indeed sometimes radical vulnerability, depending on exactly how much power they have over our lives—and we may legitimately hope that some such institutions are fit to be responsive to that dependence.

In section 1, I will present Hawley's commitment account of interpersonal trust, going on to suggest in section 2 that, *pace* her scepticism, it is virtually impossible to stop the design she tailored for individual trustworthiness spontaneously clothing collectives into the bargain, as it makes such a good fit for the institutional body. In section 3, I aim to show why it is valuable that we should possess the philosophical tools to model institutions as potentially trustworthy or distrustworthy. The main reasons I shall give are, first, that in this troubled and shifting world our efforts to make our institutions generally reliable entail efforts to render them generally trustworthy; second, that having generally trustworthy institutional bodies is of the utmost value with regard to honouring citizens' dependence upon them; and, third, that philosophically modelling institutional trustworthiness and distrustworthiness can be essential to arriving at a penetrating diagnosis of where and why a given institution may be failing, indeed betraying, the people it is supposed to serve.

[2] As reported, for instance, in *The Guardian*, upon the publication of a report on the London Metropolitan Police by His Majesty's Inspectorate of Constabulary and Fire & Rescue Services, commissioned in the wake of the the kidnapping, rape, and murder of Sarah Everard by a serving Met police officer: 'Police vetting failures have allowed 'predatory' officers to join up, watchdog finds', and 'Inspectorate for England and Wales says potentially thousands officers cleared who should have failed checks' (*The Guardian*, 2nd Nov 2022).

1. Hawley's commitment account of trustworthiness

In her 2014 paper 'Trust, Distrust and Commitment', Hawley argues that existing accounts of trust tend to ignore the phenomenon of *distrust* as opposed to a neutral attitude of mere non-trust. This, she suggests, confuses the issue, since it makes trust seem far more all-pervasive in human relations than it really is, obscuring the fact that in many situations one neither trusts nor distrusts another party, perhaps because the situation simply does not call for any attitude in that dimension—it is neutral with respect to relations of trust. Imagine a situation in which one spectator in the crowd is cheering for their friend as they run past on the track, keenly hoping the friend will be among the medal-winners in the race—they neither trust nor distrust their friend to win a medal, but simply excitedly hope that they do. Hawley's proposal is that attending fully to the phenomenon of distrust helps lend support to her commitment account of trust: trusting another person to *ø* is a matter of *believing they will fulfil a commitment to ø-ing*.

Hawley regards her commitment account as enjoying two significant advantages over the family of views she classes as 'motives-based' accounts, which tend to require some specific kind of motive for doing what one is trusted to do.[3] She argues, first, that when it comes to theorising the phenomenon of third-party trust, motives-based accounts are ill-equipped, and her own account fares much better. In a case where I, a third-party, trust an individual to fulfil a commitment to someone other than myself, Hawley's account makes ready sense of my stance of trust: simply, I believe they will fulfil the commitment. This is an attitude that is available to me even as a potentially unknown bystander towards whom the trusted party has no motivational attitudes whatever. By contrast, she argues, most motives-based accounts fail to get a grip on third-party trust, because they tend to require attitudes more typical of a second-personal case, such as presumed goodwill. Such requirements find no foothold in the situation where the person I am trusting to fulfil a commitment to someone else may not even know I exist, let alone have any particular attitudes towards me.

The second advantage Hawley argues her commitment account enjoys over motives-based accounts concerns the phenomenon of someone's being trustworthy-in-general. A motives-

[3] The accounts she discusses under this particular head are primarily Hardin (2002) and Jones' early view (1996).

based conception of general trustworthiness, she argues, would be too demanding and manipulable to boot—if it is up to others how much they invest trust in you, then you can easily become overburdened, and moreover they can burden you on purpose to suit their ends. Worse, she says, it could deliver the result that a generally distrusted person might easily achieve a near-hundred-percent track record of not letting others down, thereby counting as generally trustworthy; for if nobody trusts them, they simply have no opportunity to betray anyone. But this is hardly what it is to be generally trustworthy.

By contrast, she argues, her commitment account is especially well placed to characterize trustworthiness-in-general, casting the generally trustworthy individual as, quite simply, someone who generally fulfils their commitments. The crucial point here is that commitments, unlike others' attitudes of dependence on us, are something over which we can exercise significant normative control:

> On the commitment account, trustworthiness requires us to ensure that our commitments do not outstrip our actions. This requires judiciousness in acquiring commitments as well as doggedness in fulfilling commitments already acquired, independent of others' expectations. Trustworthy people must sometimes disappoint up-front by refusing new commitments, rather than violate trust later on: this is the moral 'power of no' (Hawley 2014, p. 15).

Hawley's commitment account, then, does not cast trust or trustworthiness as requiring any motivations in particular. She considers it sufficient for my trusting you to *ø*, that I regard you as having a commitment to *ø*, and I believe that you will *ø* —even if I do not believe that you will be motivated to *ø* by your commitment to *ø*-ing:

> According to the commitment account, you can trust someone to do something without expecting her to be motivated by her commitment. You may trust me to do something because you believe both that I have a commitment to do it, and that I will do it, without believing that I will do it *because* of my commitment. Maybe I am motivated by pure enthusiasm (Hawley 2014, p. 16).

So far as trust and trustworthiness are concerned, then, any motivation will do. However, it is important to note that this motivational promiscuity applies only to specific occasions, for regarding the determinants of a person's being trustworthy-in-general, Hawley observes, we are very unlikely to find any non-commitment-based motivations that would do the generalised motivational job. To be a generally trustworthy person, then, on Hawley's account, *will* require—if only for want of another effective general motive—a disposition to be motivated by one's commitments. That is to say, on the commitment account, the generally trustworthy individual will be motivated to fulfil their commitments for the reasons inherent in, or generated by, the commitments themselves—I will call these 'commitment-based reasons'. I would emphasize, then, that in this regard Hawley's account of trustworthiness-in-general remains importantly concerned with the motivational aspect, even while it relaxes that concern for one-off occasions.

Like most accounts, Hawley's commitment account fits well with some of our uses of 'trust' and less well with others. This is not remotely surprising. The terms 'trustworthiness' and 'trust'—like many of our most important and widely used normative concepts—are perpetually re-employed to satisfy different and evolving demands in different and newly created contexts and applications (including those in which 'trust' continues to be used for objects: 'Is this a trusted device?'). This breadth of evolving usage has the result that the word displays no single, unified proper use. A moment's reflection on ordinary usage reveals this simple fact. Imagine the following exchange regarding entrusting the cash till to someone previously caught stealing: 'But do you trust him not to put his hand in the till again?' 'Yes, absolutely!—it's okay, he just wouldn't take the risk.' And this sort of use fits nicely with Hawley's account, precisely because it does not require any particular sort of motivation from the trusted party—an extraneous motivation such as the fear of getting caught will do just fine. However, equally, the exchange might have gone like this: 'But do you trust him not to put his hand in the till again?' 'Absolutely not! Frankly, I wouldn't trust him as far as I can throw him—but it's okay, he just wouldn't take the risk.' And this sort of usage, by contrast, would not be a good fit for Hawley's account, since 'trust' is being used, with equal naturalness, in a manner that clearly does make a focused motivational demand on the potentially trusted party.

Many of our normative concepts and terms are like this—disunified in various ways, and flexible according to context and personal emphasis. Philosophical accounts effectively have

to take their pick, or split the difference, or (in my view) better, try a philosophical method that resists the professional instinct to pretend to strict definition.[4] As is so often the case, especially in theorising the ethical, if philosophy is interested in fixing a single characterization, then unless it aims to be candidly revisionist, it must offer its account in a somewhat pragmatic spirit ('Here's the most theoretically useful/common/distinctive/typical conception from our family of uses.') I read Hawley as guided by an admirable theoretical economy, so that part of what recommends her commitment account is its simplicity and ability to capture both individual occasions of interpersonal trustworthiness and also general trustworthiness into the bargain, somewhat more easily, she argues, than motives-based accounts. But my aim will not be to further engage her account of the interpersonal case. Instead, my purpose will be to address her sceptical challenge regarding the value of modelling collectives, organizations, or institutions as capable of trustworthiness as distinct from mere reliability.

2. An institutional commitment fulfilled—Already a collective trustworthiness?

Hawley's sceptical stance in relation to collective or institutional trustworthiness is, like her commitment account of the interpersonal case, driven by a certain minimalist philosophical inclination, which I share. Her chief reason for not applying trustworthiness to groups is that, she claims, we do not need to—philosophically speaking, we can rub along fine without thinking of groups and institutions in that way, and therefore do not need to stretch our philosophical resources to model any such newfangled thing. However, I would argue that hanging on to her scepticism would demand some significant negative effort, because of how very spontaneously her commitment account applies itself to institutions. Indeed, it seems to me that it is virtually impossible to stop the theory spreading to institutional bodies, so long as such bodies have commitments in Hawley's objective and motivationally undemanding sense. Once it is admitted that institutions have commitments, the commitment theory has already applied itself to such cases before we have a chance to hold it back. Moreover, I would suggest it is hard to deny that institutions sometimes have commitments. Consider a school's commitment to educate its pupils; or a corporation's commitment to prioritise the interests of its shareholders; or a government's commitment to represent the people; or a

[4] I have explored these methodological issues in relation to blaming and forgiving in Fricker 2014 and 2019.

police force's commitment to serve the public. In any given case, these commitments may or may not be fulfilled, or even properly addressed by the institutions themselves, and yet they surely exist. If believing that an agent will fulfil a commitment (through the spring of no matter what motive) is sufficient for trustworthiness in the individual case, then why restrain the spontaneous applicability of this conception to groups and institutional bodies? The fit seems irresistible, and I would have thought this a positive advantage of the account.

If we take the liberty of applying Hawley's commitment account to institutions (something she of course did not do), it is obvious that we could faithfully include its motivational promiscuity for one-off cases. If, for instance, I believe the organization that employs me will fulfil its commitment to pay its employees on time this month (perhaps I am particularly relying on this paycheque), then surely, if already persuaded by Hawley's account of interpersonal trust, I would find it hard to resist regarding the organization as *trustworthy* vis-à-vis this month's punctual pay? Matters motivational make no difference, and so in simply believing the organization will fulfil its commitment, I thereby trust. Admittedly, once we move to the idea of a *generally trustworthy* organization, things become more philosophically demanding, because here we do need to invoke *commitment-based reasons* as the motivation. This is for the reason that Hawley cites in respect of interpersonal trustworthiness-in-general: in moving from the particular to the general, we find that the only plausible motivation-type for being *generally* trustworthy is a disposition to be responsive to the commitments themselves. Thus the generally trustworthy individual is someone who fulfils the full range of their commitments for commitment-based reasons. Mirroring this now for the institutional case, the generally trustworthy institution would require an institutionalized motivational base that is effective in the fulfilment of the broadened range of institutional commitments, and for the relevant commitment-based reasons.

Modelling this sort of stabilized motivational set in an institutional body would likely call upon the mechanics of collective agency. If, for instance, we resume our example of trusting an organization to make a punctual payment this month, and extend it to exemplify an organization being generally trustworthy with respect to punctual payment, we would need to posit an institutional motive that is stable across time and counterfactual space. This temporal and modal stabilization generates something akin to an institutional *disposition*, which is what is needed to underwrite general trustworthiness regarding punctual payment. If, in addition, we extend outwards from the single issue of punctual payment to the broader class

of corporate commitments that one's employer might have (ensuring due process across its redundancy and retirement practices, serving its client base appropriately, producing an accurate annual report on time, and so on), then we would be looking at a complex *set* of institutional motivational dispositions needed to underwrite the organization's general trustworthiness *tout court*.

I have argued elsewhere that there is no significant obstacle to philosophically substantiating this idea of an institutional disposition or set of dispositions, and I have done so by reference to the idea of an *ethos*, conceived of as a set of value commitments normatively stabilized through repeated joint commitment(s) of the kind Margaret Gilbert famously uses in her accounts of collective intention, action, belief, and knowledge.[5] The joint commitments might be explicitly made to values that comprise the ethos, or they might more often be commitments to joint decisions, actions, policies, and processes that embody those values. A school board might explicitly endorse a policy whereby teachers are encouraged to be actively involved in the PTA; or alternatively it might do so more implicitly by lending support to regular PTA-organized events in the school year. Either way, these jointly committed decisions are modes of fulfilling a school commitment to be supportive of the PTA, and thereby proving trustworthy with regard to that commitment. All such commitments contribute to stabilizing the school ethos; just as joint commitments that run counter to the ethos will destabilize it. This is all the collective motivational machinery we need to model institutional trustworthiness-in-general. Granted this much, the question presents itself again, why does Hawley want to resist this natural extension by confining her commitment theory to the trustworthiness of individuals?

One reason she gives is that she suspects our ethical reactive attitudes when we demand accountability from organizations that have let us down might be more appropriately understood as directed not at the organization itself, but rather at an individual who represents it, such as a spokesperson; or, alternatively, at individuals who were in charge, or were the main actors on the day: 'although we do in fact direct trust-related reactive attitudes towards groups, with regard to their testimony as well as their actions more generally, we could more appropriately direct these towards individuals, known or unknown, group members or not,

[5] See Fricker 2010, 2012 and 2021. For an indicative sampling of Gilbert's oeuvre, see Gilbert 1989, 2004, and 2013.

who are in various ways responsible for the functioning of the group' (Hawley 2017 p. 246). This represents a theoretical option whose attractions of economy are, I hope to show, quickly outweighed by the need to hold institutions as such accountable for distrustworthiness. I won't further address this issue quite yet, however, but rather reserve it for my final section, where I will lay out an example of institutional distrustworthiness which I believe makes it plain that demands for institutional accountability at least sometimes—and perhaps in the most important sorts of case—need primarily to be directed at the relevant institutional body itself.

A more localized reason Hawley gives is that she is not convinced by the argument I put forward (specifically in relation to *testimonial* group trustworthiness) in favour of modelling the relevant epistemic institutional trustworthiness by using Gilbert's theory. In the paper that Hawley addresses, I argued that Gilbert's joint commitment conception could be applied so that the group is so constituted as to be capable of entering the relations of trust required on an assurance view of testimony.[6] The idea I advanced was that in order to make full sense of the idea of a group's testifying, we need to be able to make sense of the idea that the group in question could intelligibly *invite a hearer to trust them regarding p* (to echo Hinchman's formulation for the individual case), and that this was entirely possible if the group was in fact constituted in the manner described by Gilbert. Thus the group's members would jointly vouch for *p*'s truth, where this is achieved by way of each member (perhaps implicitly) expressing willingness to at least go along with vouching for *p* (if only by saying nothing by way of dissent as the joint commitment is being formed).

Hawley thinks this sounds all right for the 'good cases' in which well-intentioned people get together in the hope of forming a trustworthy group. As she notes, it renders intelligible both the possibility of such a group fulfilling its commitment to be trustworthy and, alternatively, the possibility that it reneges on it. Such a group might prove trustworthy in its conduct, or not; and this dual possibility is just what we wanted from the model. Let's describe such a group as *trust-apt* (short for 'trust-and-distrust-apt', rather as 'truth-apt' might be considered short for 'truth-or-falsity-apt'). However, she goes on to warn that we should also consider the radically 'bad' scenarios in which, she contends, my use of Gilbert will not work so well,

[6] See Fricker 2012, Hinchman 2005, Moran 2006, and Faulkner 2011. For an application to our relation to government see Faulkner 2018.

because, in the case she imagines, the very constitution of the group as such involves a joint commitment not to veracity but rather to mendacity, not to trustworthiness but rather to distrustworthiness:

> Imagine a nefarious group formed by members who make a mutual commitment to jointly deceive the public (insert your own example here). Such groups are constituted by the individuals' commitments to deception, not by joint commitments to trustworthiness, so do not fall under Fricker's account. Are these groups 'suited to enter into the second-personal relations of trust that characterize testimony'? Not if we construe the question literally: nobody should trust such groups (Hawley 2017, p. 242).

It is a nice point to press, especially given that there are alternative, non-assurance accounts available of group testimony, as she points out.[7] Let me first emphasize again that the view I advanced in the paper she discusses was specifically about testimony, and therefore devoted to vindicating the idea of an institution standing in the requisite relations of epistemic trust.[8] But Hawley's helpful critical point spurs me to expand the account outwards to think about how an institutional body might, more generally, be so constituted as to be trustworthy in relation to other things besides testimony, and indeed trustworthy in general, so that it might be a fitting object of a person's generalized trust. What I would like to say in reply, then, is that a group that constitutes itself, by way of any joint commitment at all, is thereby the sort of agential entity that is capable of deliberating and producing good behaviour or bad behaviour, and so capable of intelligibly serving as the object of attitudes such as trust or distrust, and many more besides. Anyone employing Gilbert's joint commitment model to craft this social and ethical possibility need not require that a trust-apt group can only acquire trust-aptitude by jointly committing *to be trustworthy* in any domain. Rather one need only observe that in becoming a group agent in the manner of joint commitment, the group thereby becomes a fitting object of such attitudes as trust or distrust, for the reason that agents as such, including group agents, are fitting objects of these attitudes.

[7] She discusses Lackey (2014) and Tollefsen (2017).
[8] The claim committed to is that 'any group partly constituted by way of a joint commitment to trustworthiness (regarding some relevant range of questions) is pre-eminently suited to enter into the second personal relations of trust that characterize testimony' (Fricker, 2012, p. 272).

Suppose, for example, we imagine a group that constitutes itself by way of a joint commitment to form a Book Club that meets weekly and puts out the occasional Instagram post about what it has been reading. Now formed, it is an up-and-running group agent, capable of all sorts of things, not just reading books and posting things on Instagram. It is capable of being, for instance, trustworthy or distrustworthy regarding what it chooses to read and how it represents its activities on Instagram. Imagine someone else had wanted to join the Book Club specifically in the hope of reading a play together, but had been told, sorry, check out our Instagram—we are only reading novels. He then learns, however, that contrary to the Instagram posts, they have in fact secretly read *Who's Afraid of Virginia Woolf?* together last month. In such a case, our disappointed would-be play-reading member will likely feel an attitude of distrust in relation to the group and how it represents its activities. If he personally knew the individuals involved, he might well feel a sting of personal betrayal, not to say hurt feelings. But the Book Club was not constituted as a group by way of a joint commitment *to be trustworthy* regarding what it read or posted; it was constituted merely through the individuals expressing willingness to function as a Book Club. That collective agency is sufficient to render the club trust-apt, along with aptness for many other attitudes besides.

My response to Hawley's point, then, is simply that agents, as such, are trust-apt; and there is no reason to consider collective agents an exception. A constitutionally distrustworthy group agent is a trust-apt agent—that is to say, an apt object of trust or distrust. To return finally to the chief reason, mentioned earlier, that Hawley resists extending her commitment account to groups and institutions: recall that she thinks we simply do not *need* it to account philosophically for our dealings with groups and institutions. After a brief discussion of Elizabeth Anderson on what it takes for an institution to be testimonially just (Anderson 2012), Hawley offers the revealing rejoinder:

> These are compelling thoughts: we need to address structural problems and collective contexts if we are to combat injustice and create better institutions. However, we do not need the distinction between trustworthiness and mere reliability at the group level in order to pursue these projects. The distinction matters at the individual level precisely because we cannot always require others to be reliable in respects which matter to us. … But we can require of our institutions that they be reliable in the respects that matter to us, or at least we can require this of our public institutions (Hawley 2017, p. 245).

Hawley is suggesting that we need relations of trust at the interpersonal level to shore up interpersonal reliability. Agreed. But she goes on to claim there is no equivalent need at the institutional level. How so? In general, as individuals, we are surely far more empowered to require of other individuals—friends, family members, colleagues, acquaintances—that they be reliable in this or that way, or even in general, because our relationships with them afford a certain accountability that bestows some authority on what we ask them to be reliable about, backed up with the usual interpersonal consequences if they let us down. The quotidian reactive attitudes apply, and they help keep us all in check on the interpersonal level. But very little of this applies in our relations with institutions, which often have immense power over us and on which we may moreover have no choice but to depend. Individuals may have some power in, say, the ability to write a scorching review online, or to officially complain, or in grievous cases to 'contest' an infringement of freedom through official channels (Pettit 1997). But beyond these sorts of mechanisms, lone individuals for the most part have remarkably little power to require anything much of institutions, for institutions are so much larger and more powerful.

Now perhaps Hawley could have agreed with me on this point, while insisting that the sense in which, as she puts it, 'we can require of our institutions that they be reliable in the respects that matter to us' was alluding to the idea that *we*, collectively and ideally speaking, can set up and regulate our institutions, especially our public institutions, as we see fit. One suspects this is what she had in mind; but the familiar god-like sense of 'require' in play here is merely theoretical and not at all what everyday social life is actually like. If, at the interpersonal level, we need to draw on relations of trust to shore up interpersonal reliability, then surely when it comes to the individual's relation to the institutions that have significant control over our lives, we need relations of trust all the more, since we are so much more dependent, indeed vulnerable, to what they do. As mentioned above, Hawley wishes to emphasize the relevance of trusting (or distrusting) individuals whose actions constitute or represent the actions of a group or institution, and she speculates that this will be sufficient to make sense of our everyday experiences of relating to groups and institutions. But I am unconvinced. Indeed I would like to argue the converse: that the value of trustworthy institutions goes to the heart of how, at least in a liberal democracy, we hope to relate to our central public

institutions, and even some of our private ones too. Let us, then, address the question: What greater value than reliability, if any, do trustworthy institutions deliver?

3. The Distinct Value of Institutional Trustworthiness

The first sort of added value that I would suggest is likely to be associated with generally trustworthy institutions, as distinct from generally reliable ones, is not a new value but rather more of the same: *more* reliability. On the defeasible assumption that, in general, motivations arising from commitment-based reasons stabilized in ethos will prove more stable (over time, in the actual world) than those arising from extraneous reasons, then an organization that deliberates from commitment-based reasons is likely to deliver more reliability than an organization that is more susceptible to the likely flux of countervailing incentives of profitability, expediency, or political favour. Other things equal, if you want a reliable institution to stay reliable in a mercurial world, you had best aim to shore up its commitments by cultivating a conducive ethos. Even this will often be insufficient, to be sure; but other things equal, it's a better bet.

Imagine a corporation that adopts green commitments at a time when such commitments are also somewhat conducive to profit, or at least compatible with it. These green policies, though not cynically embraced, are equally not connected with any stable set of values embedded in the ethos of the company. Alas we cannot on the whole predict that such policies would survive a situation in which they come into conflict with maximizing profits. Yet, by contrast, if there is a strong enough ethos in the company, some of the core values might stand a chance of outliving at least *some* degree of economic strain, or stand a better chance of resurfacing at an early opportunity. Once an institution has an ethos conducive to fulfilling its commitment set, and especially if it advertises itself as such, perhaps attracting employees or investors who care about the same things, then chances are it will more nearly fulfil those commitments over the medium and even long term.

The second value that I would argue attends institutional trustworthiness-in-general is far more distinctive, as it is famously not present in forms of mere reliability. Let us reflect, once again, on the question of motivation. Mirroring Hawley's point about trustworthiness-in-general among individuals, I have argued that generally trustworthy institutional bodies will

likely need to be fulfilling their commitments for commitment-based reasons. Now among those organizational commitments (of employers, corporations, charities, public bodies, governments, police forces, prisons, hospitals) will be commitments to the individuals and groups who *depend on them*—the employees who should be paid on time, the shareholders whose interests should be served, the clients, beneficiaries, prisoners, or patients whose needs should be met and rights respected, the disadvantaged groups whose basic welfare and dignity should be sustained, the accused who are owed due process and fair treatment. Wherever an organization has such commitments, then among its commitment-based reasons will figure the idea, inherent in any commitment as such, that the organization should not let down those individuals who are depending on them. I take it this much is built into the very idea of a commitment. A distinctive feature of some motives-based and trust-responsive accounts of interpersonal trust is their focus on exactly this special normativity of trust, which arises from the truster's dependence on the trusted party, who then (all being well) responds positively in being motivated by this *added reason* to live up to the trust placed in them. Annette Baier was perhaps the first to advance this idea, and many since have elaborated it in different interesting ways (see, for instance, Baier 1996, Jones 1996, 2012 and 2017, Faulkner 2011 and 2017, McGeer and Pettit 2017).

This question of actively depending on another person and thereby generating a special reason of trust for them not to let one down is central to the sorts of motives-based accounts from which Hawley aims to distinguish her commitment account. In one formulation, which emphasizes the dynamic, one might say proleptic, nature of placing one's trust in another person, McGeer and Pettit say that in trusting you to do something, I invest a hopeful attitude in you, thereby encouraging and empowering you to be able to fulfil the commitment with which I am entrusting you (McGeer 2008, McGeer and Pettit (2017).[9] In another formulation, Karen Jones puts it in terms of a feedback loop of 'counting on' someone to do something and their responding positively to this fact: 'This dual structure of dependency—counting on the other, in a domain, and counting on them to respond to the fact that we [are] counting on them—is the heart of trust' (Jones 2017 p. 100). In the specific context of testimonial trust, Paul Faulkner argues that an inquirer's epistemic dependence upon a testifier creates a

[9] A somewhat more instrumentally conceived version of the mechanism is explored in Pettit (1995), in which trust-responsiveness figures as 'the desire for the good opinion of others and it counts by most peoples' lights, not as a desirable feature for which they need to strive, but rather as a disposition—a neutral or even shameful disposition—that it is hard to shed' (p. 203).

distinctive reason for the testifier to be truthful, which in turn creates a bootstrapped epistemic reason to accept what they say (Faulkner 2011); and Benjamin McMyler develops the idea of a 'second-personal reason for belief' that is distinctive of testimonial trust (McMyler 2011).

In different ways, these sorts of accounts all emphasize trust's distinctive normative power by revealing the power of the relation of dependence to generate a reason to live up to the trust invested. For Jones, however, we can best understand trust by looking at it through the lens of what happens when it fails so that we are betrayed or let down—'But I was counting on you!'. These moments mark a rupture in the patterns of individuals' capacity to enlist or draw on others' agency for their own or collective purposes. In Jones' view, while it may sometimes amount to a moral betrayal, she argues that it need not, for there are non-moral kinds of value that accrue to this inter-agential facility, and which are lost when it fails. If I trust my dentist not to keep me waiting more than twenty minutes, as per their 'Commitments to Patients' pinned up in the Waiting Room, and yet I am sat in the Waiting Room with my tooth ache for an hour and a half before I am attended to, then my dentist has let me down and my trust in her to manage her schedule is undermined. But there need be nothing especially moral about this; it is plain practical, a matter of what is tolerable or convenient from my point of view. Or putting the point more strongly, as Jones notes, markedly bad people such as Mafiosi can be trustworthy for some, but that certainly does not make them or their trustworthiness morally good. Trustworthiness need not have a positive ethical valence, for it may have a negative one, or simply be ethically neutral.

At the interpersonal level, in order to isolate the distinctive functional value that relations of trust bring to our shared life, Jones imagines a fully predictable world where no one ever lets anyone else down. Perfect general reliability reigns. Yet still, she contends, something is missing that would enhance our ability to enlist each other's agency to our own or shared purposes:

> When we add trustworthiness to this world we introduce sensitivity to the reason, 'she's counting on me' and with it, the possibility of giving the one who is being counted on a direct reason to act as they are being counted on to act that did not exist prior to that very counting (Jones 2017, p. 100-101).

This thought experiment reveals a distinctive value in interpersonal trust, and in a manner that prompts me to perform the same thought experiment in relation to institutional trust. So let us take up Jones' baton here, sprinting on to institutional trust, and in particular the phenomenon of institutional trustworthiness-in-general. Imagine an ideal social world with perfectly reliable institutions, interlocking cogs turning in well-oiled organizational machines; and then imagine a trust-enhanced social world with perfectly reliable institutions but whose performances are reliable because they each have an appropriately entrenched ethos and act on motivations produced by commitment-based reasons that are firmly embedded there. This is a trust-enhanced ideal world in which our institutions are all trustworthy-in-general. The question is, Would this latter scenario exhibit any new sort of value, functional or ethical, beyond what is already present in the first? I contend the answer is Yes.

First, and once again *pace* Hawley's desire to distance her view from matters motivational, I claimed above that the very notion of having a *commitment* to someone or some group contains the idea, other things equal, that they are entitled to count on us and that we should not let them down. Just as an individual doctor, in acting on her commitment to give proper care to a patient, is thereby displaying a responsiveness to the patient's dependence on her, so is a hospital that provides care for its patients for commitment-based reasons displaying a responsiveness to their dependence on *it*. If this is right, and the very nature of fulfilling a commitment for commitment-based reasons involves dependence-responsiveness, then two things follow. First, any commitment account of trustworthiness-in-general will inevitably be drawing on the special dependence-responsive motivations of trust, because (as *per* Hawley) trustworthiness-in-general will inevitably call upon commitment-based motivations. And second, what Jones calls the 'heart of trust' is inherently in play whenever an agent acts on a commitment to another party, regardless of whether the agent is an individual or an organization.

Thus, we arrive at an institutional analogue of Jones' interpersonal point about the added normative content (dependence-responsive reasons) that relations of trust introduce into the world; and we might echo too her emphasis on the way that these trust-based reasons enable us to draw on another's agential powers to get things done. Manifestly, insofar as institutional bodies act on commitment-based reasons that involve a responsiveness to our dependence on them, our display of trust in them (perhaps just by showing up and asking for a service) is a

way of enlisting institutional agency to help us make things happen. We show up at A&E with a broken leg, and things happen; care is delivered, as best they can under stretched circumstances, because the Accident and Emergency service is committed to giving appropriate care, and because its staff respond to our manifest, acute dependence on them. If the service is trustworthy-in-general, then it reliably acts on precisely this sort of commitment-based, and dependence-responsive, motive.

In addition to this distinctive functional value of institutional trustworthiness, I believe there is also a special ethical-political value in having at least some institutions that generally act reliably from commitment-based reasons. Given that citizens have many standing dependencies on the procedures, deliberations and actions of institutional bodies—like our dependence on the PAYE tax system, or the public transport network, or health services, or the police—and, in addition, have frequent specific episodes of dependency for particular services—a tax payer applying for a rebate for overpayment, a patient seeking a referral to a specialist from their doctor, a student's having their exams marked fairly—then it stands to reason that it is valuable to us that such institutions should be responsive to these facts of dependency. If they were not so responsive, and even if they were still reliable, we would have an alienated relationship with these institutions. This is part of why we care, and are right to care, that institutional bodies at least sometimes meet their commitments to us for commitment-based reasons, and not merely because some cleverly designed incentive structure reliably produces the same results.

When an institution's policy or practice is expressive of values stemming from an ethos that we share, then it earns a valuable place in our sense of shared culture—perhaps the NHS, perhaps our employer, a green business venture, the BBC, our local school. Conversely, watching such institutions deteriorate under pressure, or become corrupted before our very eyes, is painful partly for the same reason. When institutions are motivated by commitment-based reasons, rather than reasons of expediency or profit or some other extraneous and mercurial thing, it is transforming to our relationship with those institutions. It changes the meaning of our security when we are able to trust; and it changes the meaning of our insecurity when we are not. When we trust, we relate to these institutions as *ours*—our trust is invested in them, and their dependence-responsiveness is a return investment in us. When they live up to our trust, the relation of trust in which we stand to the institution deepens the

social bond. It is entirely apt to feel betrayed when formerly trusted institutions become corrupted, for we are indeed betrayed by the institutions themselves.

Accordingly, the third and perhaps main point I wish to make regarding the distinct value of institutional trustworthiness-in-general concerns our capacity to diagnose institutionalized distrustworthiness where it may exist in our midst. If we wish to understand it, let alone fix it, then we need to be able to name the parts—to name the phenomena of institutional distrustworthiness that may be around us, or on the horizon. They will be cases of institutionalized distrustworthiness inasmuch as they depend upon the distinctive features of distrust we have discussed, such as being driven by a faulty ethos, and/or betraying reasons of trust that are generated by the relation of dependence. In the example I shall end with, both of these distinctive features are displayed in spades, so that any description of the institutional dysfunction that were to confine itself to ideas of mere unreliability, or which confined the diagnosis and calls for accountability to individual actors as opposed to directing these demands to the institution itself, would be seriously inadequate.

The example I have in mind is the long-standing and widely established (though not quite universal[10]) police interview and interrogation practice in the U.S., known as the Reid Technique. Specifically, it concerns the training program regarding the post-arrest treatment of suspects in a crime:

> [T]he Reid Technique, which is influential in the USA, encourages interviewers to use a two-stage interview process (Gudjonsson & Pearse, 2011; Inbau, Reid, Buckley, & Jayne, 2013). The first stage is a nonaccusatory interview, in which general background information about the suspect is obtained, rapport and trust are built, and a determination is made about whether or not the suspect is lying about the offence. If the suspect is judged to be lying, then the interview progresses to a nine-step accusatory (presumption of guilt) approach, typically referred to as 'interrogation' (Inbau et al., 2013). Kassin and Gudjonsson (2004) argue that the nine-step Reid

[10] For example, in 2017, having taught the Reid Technique for more than 30 years, Wicklander-Zulawski & Associates, one of the largest private police training agencies in the world, issued a press release announcing they were discontinuing the technique: https://www.prweb.com/releases/2017/03/prweb14123356.htm I thank Jennifer Lackey for this point.

> Technique can be reduced to three general phases: 'custody and isolation' (i.e. the suspect is detained and isolated, anxiety and uncertainty are generated in order to weaken resistance); 'confrontation' (i.e. the suspect's guilt is assumed and he or she is confronted with alleged incriminating evidence that may or may not be genuine, denials are rejected, even if they happen to be true, and the consequence of continued denial is emphasized); and 'minimization' (i.e. the interrogator tries to gain the suspect's trust and provides face-saving excuses for the crime, including suggesting that it was an accident or that the victim deserved it). In contrast to the PEACE Model, the Reid Technique has been associated with increased risk of false confession (Gudjonsson & Pearse, 2011; Pearse & Gudjonsson, 1999; Snook, Luther, & Barron, 2016) (Gudjonsson 2018; p. 47).[11]

This completely standard two-stage process of 'interview' followed, if suspected of lying, by 'interrogation', institutionalizes both distrustfulness and distrustworthiness in interlocking ways. The interrogators are trained, as recounted by Gudjonsson, to *presume guilt* at this stage, which means presuming the suspect is lying—an official methodological principle of institutionalized distrust. Moreover, the method relies on the idea—the still popular but empirically discredited idea—that no one would confess unless they were in fact guilty. In other work, inspired by Jennifer Lackey's philosophical analysis and the psychological research of Gisli Gudjonsson and Saul Kassin, I have dubbed this discredited idea The Confession Myth (Fricker, forthcoming; Lackey 2020 and forthcoming; Gisli Gudjonsson 2018; Kassin 2004, 2010, 2022). In Lackey's revealing analysis she emphasizes the central role of what she calls Agential Testimonial Injustice, defined as a credibility excess given only when someone has been epistemically broken down, perhaps so badly that they might say almost anything. So it is in cases of extracted confessions. In my own treatment I have emphasized how we need to attend to a three-phase pattern of, first, institutionalized prejudicial credibility deficit (Testimonial Injustice) when pleas of innocence are prejudicially rejected; followed by credibility excess (Lackey's Agential Testimonial Injustice) when a confession from an epistemically disempowered subject is secured; and

[11] The PEACE model [Preparation and planning, Engage and explain, Account, Closure, and Evaluation] was developed in the UK in 1993 after a number of shocking miscarriages of justice were revealed. The technique was later adopted in New Zealand and Norway. Gudjonsson describes it as having been 'developed on sound psychological principles, which followed valuable collaborative work between academics, psychologists, police practitioners, and lawyers. It was intended to take into account the vulnerability of some interviewees, with the aim of minimizing the risk of a false confession' (Gudjonsson 2018, p. 47).

finally, a return to credibility deficit again, at the procedural point where rescinding one's confession is rendered formally impossible so that the confessor's renewed pleas of innocence are doomed to meet with the final phase of institutionalized distrust.

Under the Reid Technique, the aim of interrogation is to get the suspect to incriminate themselves, ideally signing a confession, and since the whole process officially and explicitly institutionalizes the presumption of guilt, the police are legally entitled to lie and epistemically manipulate in order to secure whatever self-incrimination may result, which in turn will be taken as proof positive of the guilt that is presumed from the start. In Saul Kassin's recent book, consolidating over two decades of work on false confession, *Duped: Why Innocent People Confess—and Why We Believe Their Confessions*, he documents and explains many cases of false confession. With the tools available to the police, operating according to the Reid Technique, one of the things that actually makes it more likely that a suspect winds up confessing is *innocence*. One key explanatory factor here is that an innocent suspect is more likely to waive their rights to an attorney and to silence. This, and their willingness to disclose details, since they feel they have nothing to hide and wish to demonstrate helpfulness to the police, renders them lambs to the slaughter when it comes to the standard techniques the interrogators are trained to use.

If the suspect is a minor, then the vulnerability to emotional and epistemic manipulation becomes even greater, even before we reflect on the trauma of being held in custody and subject to racism. An infamous case is that of the five black teenagers falsely accused of the rape and beating of a white woman who was jogging in New York's Central Park in 1989.[12] As Kassin et al describe:

> Each boy retracted his statement immediately upon arrest, saying he had confessed because he expected to go home afterward. All the boys were convicted and sent to prison only to be exonerated in 2002 when the real rapist gave a confession, accurately detailed, that was confirmed by DNA evidence (Kassin et al 2010, p. 4).

[12] The award-winning dramatization, 'When They See Us', was created and directed by Ava DuVernay in 2019 (Netflix limited series). See also Ken Burns's 2012 documentary 'The Central Park Five'.

The practice of police interrogators lying even to children deemed suspected of a crime remains perfectly legal in 48 of the 50 states—Oregon and, more recently, Illinois being the only two exceptions at the time of writing.[13]

The operation of the Reid Technique, with its built-in presumption of guilt and methods of deception and manipulation, displays a powerful form of institutionalized distrust in the presumption of guilt. And what the psychology of the past two decades has taught us is that the whole process is to be *distrusted*. It is not reliable, but in addition it would be implausible in the extreme to think that the institutionalization of lying and manipulating suspects under interrogation creates anything other than a bad ethos in relation to the treatment of suspects, and redirects what should be a process of genuine inquiry so that it is instead a process of actively constructing institutional 'proof' of guilt presumed on the basis of the interview. The result is therefore not merely a problem of unreliability—though it certainly is that—for the police are failing to deliver on the institutional commitment to conducting a proper investigation. The dysfunction calls for capture in the idea of *institutionalized distrustworthiness*.

That we need to focus philosophically on the trust dimension in this case is further evidenced by the fact that the other distinctive mark of trust and the betrayal of dependence is also present. The implementation of the Reid Technique actively exploits suspects' vulnerabilities in general—including youth, emotional stress, or mental impairment—but in particular it exploits the fact that many people are not aware that the police are legally entitled to lie to them in efforts to lead them to incriminate themselves under interrogation, rather than trying without prejudice to gather more evidence one way or the other. This creates a situation in which suspects are *counting on* the police not to mislead them about the legal process, or the consequences of staying silent, or of continuing to profess their innocence…so that when the police in fact lie about these things in order to get them to incriminate themselves, the suspect's manifest dependence on them is betrayed. The result can too often be that when a teenager is told that if he just signs on the dotted line then he can go home to his parents, he believes them; and if he signs, there will be little besides the holy grail of an after-the-fact

[13] Currently, in New York state, Senate Bill 324, sponsored by Zellnor Myrie, proposes the banning of police deception in interrogation and would require that courts assess the reliability of confessions prior to their being admissible. It is still in committee—NY awaits its progress through the legislative process.

DNA exoneration that could reverse the legal power of the confession to constitute him as institutionally guilty.

As I mentioned earlier, Hawley proposes that we can and should direct our natural reactive attitudes of accountability towards individuals acting under the auspices of the institution rather than towards the organization itself. This is part of her strategy for showing that we can theorise everything we need to theorise without getting embroiled in talk of trusting institutions or groups. But in our present example I believe we see that this is not so. While it is entirely open for someone to direct blame and related reactive attitudes towards individual police officers who manipulate suspects as they are trained to do according to the Reid Technique, still to do this *instead of* demanding accountability from the institution itself would be to miss the fact that the practice of betrayal is procedural, and fully institutionalized. The presumption of guilt and the use of tools of manipulation are all institutional procedures enshrined in police practice and the U.S. law. This suggests that our outraged disillusionment and blame for these processes should be directed first and foremost at the relevant institutional agents—the institutional bodies, organizations and corporations that create, sustain, and profit from them.

This ends my case for the distinctive value afforded by having institutions that are trust-apt as opposed to merely reliance-apt. The example of a fully institutionalized case of distrustworthiness is salutary. I hope I have shown Hawley's challenge to be an excellent heuristic. We should indeed ask, What added value, if any, does it bring to think in terms of group trustworthiness or distrustworthiness? I have made a case for an answer that has emphasized, first, the sheer *inevitability* of thinking in terms of group trustworthiness as soon as we adopt a commitment account. I have gone on to argue that we do need to theorize trustworthiness in organizations, for there are some institutional bodies and processes whose dysfunctionality and moral status we can only fully understand if we pay attention to *ethos*, and the potential betrayal of individuals and groups that depend on them. And lastly, I have presented a contemporary example from the widely recommended police interrogation method in the U.S., which shows that distrustworthiness at least sometimes inheres in the institution itself by dint of the official processes and policies it imposes on those who are radically dependent upon its authority. If we want to understand our social world and

accurately diagnose its dysfunctions, I conclude we do need to think in the enriched philosophical terms of institutional trustworthiness and distrustworthiness.[14]

References

Anderson, Elizabeth (2012) 'Epistemic Justice as a Virtue of Social Institutions', *Social Epistemology* Special Issue on Epistemic Injustice 26(2): 163-73

Baier, Annette (1996) 'Trust and Antitrust', *Moral Prejudices: Essays on Ethics* (Cambridge, MA: Harvard University Press)

Faulkner, Paul (2011) *Knowledge On Trust* (Oxford: Oxford University Press)

Faulkner, Paul (2017) 'The Problem of Trust' in *The Philosophy of Trust*, eds. Paul Faulkner and Thomas Simpson (Oxford: Oxford University Press)

Faulkner, Paul (2018) 'Finding Trust in Government, *Journal of Social Philosophy* 49(4): 626-644

Fricker, Miranda (forthcoming) 'Institutionalized Testimonial Injustices: The construction of a Confession Myth' *Journal of Dialectics of Nature*, Special Issue on Epistemic Injustice ed. Huiren Bai

Fricker, Miranda (2021) 'Institutional Epistemic Vices: The Case of Inferential Inertia' in Heather Battaly, Quassim Cassam & Ian J. Kidd eds. *Epistemic Vice* (London: Routledge)

Fricker, Miranda (2019) Forgiveness: An Ordered Pluralism', *Australasian Philosophical Review* 3/3 (2019): 241-260

[14] I am most grateful to an anonymous reviewer for helpful comments on an earlier draft.

Fricker, Miranda (2014) 'What's the Point of Blame? A Paradigm Based Explanation', *Noûs* (2014) 50 (1): 165-183

Fricker, Miranda (2012) 'Group Testimony? The Making of A Collective Good Informant', *Philosophy and Phenomenological Research* (84) 2: 249-276

Fricker, Miranda (2010) 'Can There Be Institutional Virtues?' in Tamar Gendler & John Hawthorne eds. *Oxford Studies in Epistemology* Vol 3: 235-252

Gilbert, Margaret (1989) *On Social Facts* (Princeton NY: Princeton University Press)

Gilbert, Margaret (2004) 'Collective Epistemology' *Episteme* 1(2): 95-107

Gilbert, Margaret (2013) *Joint Commitment: How We Make the Social* World (Oxford: Oxford University Press)

Gudjonsson, G. H. & Pearse, J. (2011) 'Suspect interviews and false confessions' *Current Directions in Psychological Science*, 20: 33-37

Gudjonsson, Gisli H. (2018) *The Psychology of Confessions: Forty Years of Science and Practice* (Chichester UK: John Wiley & Sons)

Hardin, Russell (2002) 'The Reasons of Trust' *Australasian Journal of Philosophy* 86(2): 213-36

Hawley, Katherine (2017) 'Trustworthy Groups and Organizations', essay 14 in *The Philosophy of Trust*, eds. Paul Faulkner and Thomas Simpson (OUP)

Hawley, Katherine (2014) 'Trust, Distrust and Commitment' *Noûs* 48(1):1-20

Hawley, Katherine (2012) *Trust: A Very Short Introduction* (Oxford: Oxford University Press)

Hinchman, Edward (2005) 'Telling as Inviting to Trust', *Philosophy and Phenomenological Research* Vol. LXX, No. 3: 562-87

Inbau, F. E., Reid, J. E., Buckley, J. P., & Jayne, B. C. (2013) *Criminal interrogation and confessions* (5th ed.) (Burlington: Jones and Bartlett Learning)

Jones, Karen (2017) 'But I Was Counting on You!' in *The Philosophy of Trust*, eds. Paul Faulkner and Thomas Simpson (Oxford: Oxford University Press)

Jones, Karen (2012) 'Trustworthiness' *Ethics* 123(1): 61-85

Jones, Karen (1996) 'Trust As An Affective Attitude' *Ethics* 107: 4-15

Kassin, S. M. (2022) *Duped: Why Innocent People Confess—and Why We Believe Their Confessions* (Amherst/New York: Prometheus Books)

Kassin, S. M., Drizin, Steven A., Grisso, Thomas, Gudjonsson, Gisli H., Leo, Richard A., Redlich, Allison D., (2010) 'Police-Induced Confessions: Risk Factors and Recommendations' *Law Hum Behav* 34: 3-38

Kassin, S. M., & Gudjonsson, G. H. (2004) 'The psychology of confessions. A review of the literature and issues' *Psychological Science in the Public Interest* 5: 33-67

Lackey, Jennifer (2014) 'A Deflationary Account of Group Testimony' in Lackey (ed.) *Essays in Collective Epistemology* (Oxford: Oxford University Press)

Lackey, Jennifer (2020) 'False Confessions and Testimonial Injustice' *Journal of Criminal Law and Criminology* 110 (1): 43-68

Lackey, Jennifer (forthcoming) *Criminal Testimonial Injustice*

McGeer, Victoria (2008) 'Trust, hope and empowerment' *Australasian Journal of Philosophy* 86(2): 237-254

McGeer, Victoria & Pettit, Philip (2017) 'The Empowering Theory of Trust' in *The Philosophy of Trust*, eds. Paul Faulkner and Thomas Simpson (Oxford: Oxford University Press)

McMyler, Benjamin (2011) *Testimony, Trust, and Authority* (Oxford: Oxford University Press)

Moran, Richard (2006) 'Getting Told and Being Believed', in J. Lackey and E. Sosa eds. *The Epistemology of Testimony* (Oxford: Oxford University Press)

Pearse, J. & Gudjonsson, G. H. (1999) 'Measuring influential police interviewing tactics: A factor analytic approach' *Legal and Criminological Psychology* 4: 221-238

Pettit, Philip (1997) *Republicanism: A Theory of Freedom and Government* (Oxford: Oxford University Press)

Pettit, Philip (1995) 'The Cunning of Trust' *Philosophy and Public Affairs* 24(3): 202-225

Snook, B., Luther, K., & Barron, T. (2016) 'Interviewing suspects in Canada' in D. Walsh, G. E. Oxburgh, A. D. Redlich, & T. Mykelburts (eds.) *International developments and practices in investigative interviewing and interrogation* Volume 2: Suspects (pp. 229-239) (London: Routledge)

Tollefsen, Deborah (2007) 'Group Testimony' *Social Epistemology* 21.3: 299–311