From Trust to Trustworthiness

Philosophers have written a lot about trust, but we have been surprisingly silent about trustworthiness. There are scattered remarks about it in discussions of trust, so the silence is not complete, but the problem of understanding trust and trustworthiness has been pursued largely from the trust side.[1] Despite this silence, we can read what philosophers must have been thinking about trustworthiness from what they have said about trust, since accounts of trust typically contain reflections of an implicit account of trustworthiness, glimpsed backwards like writing viewed in a mirror. Accounts of trust have natural partners in accounts of trustworthiness, and not just any pairing will be a good fit. Thin accounts of trust, such as risk-assessment accounts, which analyse trust as a subjective degree of probability that the one-trusted will perform the desired action, fit with thin accounts of trustworthiness that demand only that the trustworthy have motives sufficient to get the job done. Thicker accounts of trust require correspondingly thicker accounts of trustworthiness.

There is now a whole menu of accounts of trust to chose from: from the thinnest that equate trust with reliance, whether confident or not, to the thickest that suppose it to require a belief in the good will or integrity of the one-trusted. Some accounts emphasize affect, some belief, some action, others combinations of the three.[2] Nor is there (yet) any sign of convergence. Philosophers have used

the method of narrow reflective equilibrium to defend their preferred conceptions of trust. What account best explains and brings into coherence our various intuitions about trust, including intuitions about particular cases – is it, for example, trust or just reliance if you depend on the other's fear? Convergence has proved elusive, however, because intuitions differ and we lack a theory as to why some should have the status of "considered intuitions" and so be taken seriously in our theory-building while others can safely be ignored.

When faced with the failure of a method to bring convergence, it's time to back up and ask, have we got our method right? Could we have been missing some of the constraints that a good account of trust must meet and so failing to get convergence? And what could those extra constraints be? To answer that, we have to back up even further and ask what work we want the concept of trust to do, for only when we understand what the point of a concept is, can we tell if a proposed account of it is any good.

This article intertwines both substantive and methodological goals. Its substantive goal is to defend an account of trustworthiness, but to do this it is necessary to move beyond the methodology of intuition juggling. Trustworthiness is interesting in its own right, but part of the interest in focusing on it comes from the light it promises to shed on outstanding questions in the philosophy of trust. Philosophers' comparative silence on trustworthiness is part of the explanation for our failure of convergence on trust. If, as I argue we should, we think of trust and

trustworthiness as paired concepts that are to be investigated in tandem, with each setting constraints on the correct understanding of the other, then we can find the additional constraints we need to help resolve some outstanding disagreements in the philosophy of trust, including, most especially, disagreement over the motivational structure trust imputes to the one-trusted. We find the extra constraints by first examining the point of having these paired concepts.

Once we switch perspective and approach the problem of understanding the pair from the trustworthiness end, their normative role comes more clearly into view. Trustworthiness and trust are not reducible to reliability and reliance because they identify, in order to promote, a distinctive way that our cognitive sophistication make it possible for us to respond to the fact of interpersonal dependency. I argue that this role is best served by an account of trustworthiness as competence together with direct responsiveness to the fact that the other is counting on you. So understood, "trustworthiness" names something less than a virtue, but it nonetheless identifies a source of motivation that it is of vital interest to finite social beings, such as ourselves, who have everything to gain – or to lose – from engaging in relationships of dependency.

1. Getting started: Two methodological principles

I begin by defending two methodological principles that will guide my investigation. Together, the principles are powerful tools to use in generating and evaluating accounts of trust and trustworthiness. Nevertheless, despite their power, they deserve wide acceptance. That they have not secured such acceptance is due, I think, to a general lack of reflectiveness about methodology and to confusion about what accepting them would entail.

1.1 Trust and trustworthiness are paired concepts that need to be investigated in tandem.

Our theory of trust must mesh with our theory of trustworthiness and vice versa because the two stand to each other in the normative relation of "fit". Trust is a fitting response to trustworthiness and the trustworthy are fit objects for our trust. On the face of it, this assumption seems obvious enough. It is implicitly recognized in the literature; for example, both Annette Baier and Karen Jones recognize it when, in generating their accounts of trust, they help themselves – without explanation or defence – to constraints that come from thinking about trustworthiness, a strategy legitimate only on the assumption that we are dealing with paired concepts that must somehow "fit together".[3] It is thus somewhat puzzling that this strategy has been neither defended nor systematically pursued. The explanation must surely lie in concerns about what else you would be

committed to if you saw trust and trustworthiness as closely paired, mutually constraining, concepts. Russel Hardin, who explicitly demands that trust and trustworthiness be understood in tandem, also buys into a controversial package of claims about when trust is justified.

Hardin holds that to trust just is to believe that the other is trustworthy, and (I'll return to say more about this in section 3.2) to be trustworthy with respect to a person and an action is for it to be in your interest to act in their interests in this matter.[4] Hardin also thinks that trust cannot be justified on instrumental grounds and cannot be rational unless based on evidence that the other is trustworthy.[5] Trust succeeds – which on this view amounts to "is true" – if *and only if* the one-trusted is trustworthy. To trust the untrustworthy is necessarily to make a mistake, for only the trustworthy merit trust; to extend trust on grounds other than evidence of trustworthiness is also to make a mistake, of the same kind Pascal makes famous in his wager.

Suppose that, unlike Hardin, we wanted to keep open the possibility that trust can be justified on the basis of forward-looking considerations, that we might knowingly and optimistically extend trust to those who we very much doubt are trustworthy, in the hope of thereby eliciting trustworthiness from them. If trust counts as correct when it is a response to the trustworthy and if it is trustworthiness that merits trusts, then wouldn't optimistic trust have to count as mistaken and hence, whenever it is knowingly undertaken, as unjustified and

irrational? It might appear that the alleged truism that trust and trustworthiness stand in the normative relation of "fittingness" to each other is no truism at all, but a way of smuggling in hard headed – and hard hearted – assumptions about when trust could be justified.

The inference does not follow. There is an analogy between the relationship that trust holds to trustworthiness and the way in which an emotional attitude can be, or fail to be, a fitting response to the properties of an evoking situation. Fear, for example, is a fitting response to the dangerous, shame to the shameful; and, I claim, trust to the trustworthy. Building on the analogy with the relationship between affective attitudes and evaluative properties, we can say that there is a sense in which the attitude *succeeds*, and so counts as correct, when it represents the evoking situation as having an evaluative property it in fact has.[6] Thus trust succeeds when it is directed at the trustworthy – it responds to them as having a property that they in fact have. This is what prevents us from pairing a thick account of trust – trust as optimism about the good will and competence of the other, say – with a thin account of trustworthiness, as requiring only psychological states sufficient to get the job done. But from the fact that trust succeeds when directed towards the trustworthy since it represents them as having a property they in fact have, it does *not* follow that trust that is directed at the trustworthy is all-things-considered justified. Suppose we have neither reason to believe that the person is trustworthy, nor reason to think that our trusting would

elicit trustworthiness from them, nor some other good reason for extending trust, then trust would be unjustified, even though it would succeed in the limited sense of responding to the person as having a property that they in fact have.

A normative relation holds in the reverse direction as well, but again nothing untoward follows from this. Consider once more the analogy with affective states and their objects. To be shameful is to be such as to *merit* shame; to be trustworthy is to be such as to *merit* trust.[7] It is in trust that trustworthiness receives its proper uptake and recognition. This explains why a thick account of trustworthiness should not be paired with a thin account of trust: if trustworthiness requires integrity, say, but trust only requires an expectation of a positive outcome, then trust will not recognize trustworthiness, but rather something lesser, such as reliability. Nothing that's been said so far implies that we cannot be justified in trusting the untrustworthy, nor that they could not deserve trust, on moral grounds, say. Perhaps we can, perhaps we can't. I mean to remain agnostic about these questions, here. The claim that matters – that underwrites the method of seeing trust and trustworthiness as paired concepts that must be investigated in tandem and that normatively constrain each other – is the weaker claim that trust is a fitting response to trustworthiness and that trustworthiness merits trust. Shorn of assumed implications that do not in fact follow, it should receive general acceptance and it is enough to underwrite a methodology of working the problem from now one end, now the other.

1.2 Trust and trustworthiness must earn their keep: they are concepts with both normative and explanatory roles

Good concepts do useful conceptual work. That's what prevents us from just introducing a word and stipulating that it names a heterogeneous bunch of things that share nothing interesting in common – unless, of course, we are in a lecture and making a philosophical point about the silliness of doing so. The point of many concepts is to identify a class of entities that have the kind of unity needed to sustain interesting generalizations and to play explanatory roles in our best theories of the domain in question.[8] The social science literature recognizes that "trust" should identify a category with enough unity to sustain generalizations and play an explanatory role: common explanatory claims include that trust explains (at least some) cooperation, that trust lowers transaction costs, and that high trust societies present more social and economic opportunities than low-trust societies.[9] If an account of trust identifies too heterogenous a class of dependencies to do such explanatory work, as is the case with accounts that reduce trust to reliance, with or without confidence, then it is inadequate.[10]

Other concepts have normative roles: moral concepts such as ought, right and the various virtue terms, as well as epistemic concepts like warrant and justified are obvious examples. There is recognition that "trust" might have a normative role in moral psychology work on the notion, in accounts that align it with our practices of holding each other responsible through reactive attitudes, or

in accounts that align trustworthiness with goodwill, or integrity.[11] However, though defended by some philosophers,[12] there has been a push away from moralized conceptions of trust and trustworthiness because trust can enable exploitation and abuse just as much as it can enable cooperation. Trustworthiness can be in the service of bad ends as easily as good ones and sometimes we can be morally required to respond to trust with judicious "trust-busting".[13] If trust and trustworthiness do not name virtues, it might be inferred that their role cannot be normative and so, if they are to have any useful role at all, it must be for those purposes of explanation and induction emphasized by social science.

The inference does not follow: a concept can have *both* explanatory *and* prescriptive roles and yet not be a moralized notion at all. "Human kinds" have these dual roles:

> Many human kinds have powers unknown to natural kinds. They are instruments and agents of power and knowledge, but also of caring and of stewardship. … I think that the role of human kinds in our lives, and in the human and social sciences too, has little to do with those spectator sports so admired in some theories of natural kinds, namely induction and explanation.[14]

We understand such categories when we understand their normative point. The availability of these concepts itself contributes, via a looping mechanism, to

bringing about the causal regularities that sustain generalizations about the members of the class that fall in their extension and enable the terms to have an explanatory role.[15] Consider the gender-role terms "masculine" and "feminine". At least some of the generalizations that can be made about men and women come about because we accept the terms "masculine" and "feminine" and make our behaviour confirm to the norms that they implicitly embed.[16] In section 3.3, I'll argue that the availability of the concept trustworthiness enables just such bootstrapping, further supporting the contention that it and its partner, trust, are human kinds.

The example of gender-role terms, which have both explanatory and prescriptive functions, demonstrates that there is conceptual room for the paired concepts of trust and trustworthiness to have a normative role without either of them naming virtues. Once we have a clearer conception of what trustworthiness is, I'll return to the question of whether it is a virtue; for now it is sufficient to note the we can accept that they have both normative and explanatory roles while remaining agnostic about whether they name virtues. Thus, if your reason for overlooking possible normative roles for trust/trustworthiness has been either that you thought you had to choose between normative and explanatory, or that normative entailed moralized, then you need have no concern with accepting my "both/and" methodological principle. Once alive to this possibility, it becomes

easier to see why we might bother having these paired concepts, just what conceptual work they might do for us.

2. **An hypothesis about the role of the concepts trust and trustworthiness**

Why do we need a distinctive pair of concepts tailored to apply (in their core uses) to fellow human beings? Why not make do with the perfectly general concepts of reliance and its twin reliability, which can apply equally to human and to non-human agents as well as things? In this section, I offer a conceptual "job-description" for trust and trustworthiness, asking not what our concepts are, but what – in broad outline – we should want them to be able to capture, if they are to do useful conceptual work. Conceptual role arguments have the potential to be revisionary: we could find out that the concepts we in fact have are but poor candidates for the given job-description and that they need to be replaced wholesale, or at least substantially revised.[17] But, as will become apparent in Section 3.2, I think that the conceptual role argument is able to generate an account of trustworthiness that can explain what is right about most of the serious contenders for a conceptual analysis of trustworthiness that can be distilled from the literature. Thus the account we arrive at by beginning from the question "what should our concepts be?" is not wholesale different from the one we could have arrived at beginning from the question "what are our concepts?". There is a good reason our concepts are (roughly) as they are.

Trust and trustworthiness have work to do because of three fundamental facts of human existence: we are finite, reflective, and social creatures. As social beings, other agents are a particularly salient source of risk to us, but they also provide a remedy for our finitude, for together we can do what neither of us can do alone. Moreover, we do not stand to each other in the same relation that forces in the natural world (rivers, say) stand to us. We can moderate the risks we face from natural forces by taking measures to control them: canals and embankments will reduce the risk of flooding. However, the measures we can take to control mere natural forces are necessarily limited in the sophistication of their interactivity. A natural force can be controlled causally and it can causally effect what we will do, but it has neither the capacity to control its own behaviour, nor the capacity to anticipate our behaviour. Only agents, whether human or animal, have the ability to control and to anticipate.

Animal agents have the capacity to modify their behaviour in the light of their anticipation of the behaviour of others and so can pursue behavioural strategies that embed assumptions about what other creatures will do. Dogs do this when they respond to other dogs signalling their intention to attack, to submit, or to play. In this sense, their behaviour can depend on their understanding of the behaviour of others and so displays a first level of interactive sophistication.

Humans have the sophistication to do more than this: we have the cognitive capacity to take into account in our deliberation the fact that another

agent's deliberation rests on assumptions about what we will do. This capacity requires not just a mind and the capacity to make decisions, but a *theory of mind* and the capacity to make decisions taking into account the mental life of the other, including their beliefs, intentions, desires and expectations.[18] And it opens up the possibility of explicitly taking into account the fact that others are counting on you.

To count on something or someone is to embed in your plans an expectation that, if false, means you will be left worse off than you otherwise would have been. The success of your action or plan depends, non-trivially, on what you are counting on coming to pass. 'To count on' is thus to something more than merely to expect: we have all sorts of expectations about the behaviour of things and people, but only some of them come to be embedded in our plans in ways that affect their success. These are the things we count on, whether consciously or not. In many cases, it is our counting that creates a dependency: the success of our action is now dependent on, or hostage to, the behaviour of the thing or person that we count on. Had we recognized the assumption embedded in our plans and thought it unlikely to be fulfilled, we would not have gone ahead. There can also be inescapable dependencies, where we have no choice but to count on an outcome even in the face of grave doubt.

Our ability to take into account in our deliberation the fact that others are counting us makes available to us a distinctive way of responding to the fact of

other agents' dependency *through recognizing that very dependency*. Nor do the implications of our cognitive sophistication end there. We each know that the other – provided they have reached a sufficient level of maturity – is able to take into account the ways in which the success of our action depends on what they will do. This opens up another level of sophistication in the interactivity possible in managing our dependencies: we can count on the other responding to our counting on them. That is to say, we can embed in our plans the assumption that the other will recognize and respond to the fact of our dependency. And they likewise can recognize this, and can respond to this new way in which they are being counted on and may do so even when they would not have responded to first level dependency.[19]

Knowing that others can themselves recognize and respond to our dependency means we can actively seek to recruit their agency to enhance the effectiveness our own. Of course we do not always do this. Sometimes we treat other agents much like natural forces whose behaviour is a mere regularity to be worked with or gotten around. At other times we recognize that what they will do depends on what they think we will do, where they also recognize that we recognize this very fact, and so we anticipate each other's behaviour but in a context in which we are each going about our own business. Share trading illustrates this kind of complex interaction of expectations among agents each going about their own business (literally). I think that everybody else will think

everybody else will sell and I decide to ride out the expected slump, or follow the trend depending on my debt exposure. In these and a myriad of like cases, we depend on each other in the sense that the success of our action is vulnerable to the other's choice of action, and we recognize this, but we do not depend on the other responding to that dependency. When we add this extra level of reflectiveness about our dependency we get into a territory where there is some distinctive work to be done that cannot be done by the concepts of reliance and reliability, which apply to forces of nature and non-reflective agents as well as to fellow human beings. I suggest that this is the distinctive work for which we need our concepts of trust and trustworthiness.

Trust and trustworthiness are concepts that bring into focus inter-human dependencies and draw our attention to the special capacity for responsiveness to those dependencies that our reflectiveness makes possible. We have a use for the twin concepts of trust and trustworthiness to mark the distinctive way our cognitive sophistication makes it possible for us to respond to our vulnerability at the hands of other agents through active engagement with their agency and they can respond to the power of their actions to bring us good or ill through active engagement with the fact of our dependency.

The purpose of the concepts is broadly normative: we focus on this distinctive kind of active dependency and active responsiveness to it so as to promote them as ways of extending our agency through dependency on others.

This is borne out in moral education, where a child's attention is specifically drawn to the fact that others are counting on them and to the possibility of their living up to others' expectations in the hope of thereby fostering such responsiveness. It is also borne out by trust-responsiveness: sometimes displaying trust is sufficient to elicit trustworthiness as we respond to the call to be moved by the other's dependency.

None of this is yet to say exactly *how* trust and trustworthiness are, from opposite sides, ways of actively engaging with the fact of human dependency. But if this story is along the right lines, then accounts of trustworthiness are to be evaluated according as they enable us to cash out the thought that trustworthiness is a way of actively and positively engaging with the fact of the other's dependency, made possible by our capacity to recognize such dependency and to take it into account in our deliberation. As finite and social agents, we have a pressing interest in ways of recruiting our reflective abilities to reduce the risk of dependency. We are thus interested in labelling that distinctive mode (whatever exactly it is) of engaging with the other's dependency in order to promote it.[20]

3. Trustworthiness

3.1 Against Thin Accounts

We are able to use the principle that accounts of trust contain mirror-image accounts of trustworthiness to overcome philosophers' comparative silence on trustworthiness and extract implicit theories of trustworthiness from explicit theories of trust. A "get the job done" account of trustworthiness fits with risk assessment accounts of trust. And we have the resources to make quick work of it (and, by application of Principle 1, its corresponding family of accounts of trust): it identifies a notion that can do no work at all, neither explanatory, nor normative.

Gambetta provides an example:

> Trust (or, symmetrically, distrust) is a certain level of subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both *before* he can monitor such action (or independently of his capacity ever to be able to monitor it) *and* in a context in which it affects *his own* action…When we say that… someone is trustworthy, we implicitly mean that the probability that he will perform an action that is beneficial to us or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation.[21]

Risk-assessment accounts do not distinguish among the reasons why the truster assigns a sufficiently high degree of probability to the one-trusted performing the action. They count as trustworthy with respect to an action Z, any agent who has psychological structures sufficient to get the job done. This minimalist conception of trustworthiness makes it a two-place relation: "B is trustworthy with respect to Z." The truster drops out of the picture and trustworthiness does not specify an interpersonal relation but rather a relation between a person and a deed. It thus follows that it cannot be a way of cashing out how the trustworthy engage with the dependency of others, for each party might just be going about their business without any engagement at all.

Any motive sufficient to perform Z will do, be it fear, habit, vanity, virtue, self-interest, or goodwill. It follows from this that the category "trustworthy" or even "trustworthy with respect to Z" sustains no useful inductive generalizations and will support no explanations, except the trivial "apt to do Z" which follows from the definition itself. The notion is explanatorily idle. Nor can it play a role in identifying a grouping for purposes of the normative roles of "caring and stewardship". There being nothing that unites the trustworthy, not even the trustworthy with respect to an action Z, there can be no theory of how to promote trustworthiness, nor indeed any reason to think trustworthiness is, other things being equal, something that is to be promoted. If we want our category

"trustworthy" to do any useful work, whether explanatory or normative, or both, then we cannot draw it so broadly – a concept so broadly drawn is useless.[22]

3.2 Trustworthiness as active engagement with dependency

Suppose we take the seriously the suggestion that the role of the concept of trustworthiness is to identify, in order to promote, a distinctive way in which human beings can actively and positively engage with the fact of another's dependency, through their ability to recognize it. The conceptual role argument brings into focus the pressing interest that we have in there being people who will *take* the fact that others are counting on them to be reason-giving in their practical deliberation, but it is neutral over the question as to whether there must be a deeper story as to why they do this. In this section, I show that what unifies the various positive states that have been put forward as possible motives for the trustworthy is that there are contexts in which they explain why an agent would take the consideration "he's counting on me" to be reason-giving. In contrast, what unifies the various states taken to be trustworthiness undermining is that they cannot play this role. Thus, we have a simpler theory if we just cut to the chase: the trustworthy (with respect to a person in a context) take the fact of someone's depending on them to be reason-giving. There may be a deeper a story as to why this is so, but there does not have to be, and it need not always be the same story.

Before the argument can get off the ground, we need to get clearer about what it is for an agent to take a consideration as reason-giving. As a rough gloss, a practical reason is a consideration that counts for or against an intention or course of action.[23] They include considerations such as "it would be fun," "he needs my help," and "I promised." An agent takes a consideration to be reason-giving if and only if she accords it a justificatory role in her practical deliberation. Justificatory roles can include taking it weigh in favor or against, or taking it to defeat or modify the justificatory force of other considerations. "Taking as a reason" is thus a notion in moral psychology, rather than in normative theory: one can take to be reasons considerations that are no real reasons at all; one can even take a consideration to be reason-giving without judging that it is.[24]

An agent typically takes a consideration to be reason-giving because of background facts that explain her responsiveness to that consideration. Call these, "background reason-enabling conditions." If a condition C is reason-enabling for a consideration R, then when C is present the agent will tend to track R and to assign it a reason-giving role in her deliberation. She may not invariably be responsive to R: fatigue and lack of time to deliberate can explain failures in responsiveness, as can explicitly engaging self-regulative capacities to override the effects of the background reason-enabling condition when the agent judges she should not take a consideration to be reason-giving.

Included among background reason-enabling conditions are states that are sometimes subject to self-regulating override such as character traits and occurrent emotional states, as well as those that do not attract such regulative override, such as values, norms that the agent accepts regarding what considerations to treat as reason-giving and, especially in shared deliberation, decisions and agreements about what considerations are to be assigned a justificatory role.[25] For example, "he needs my help" is a consideration that can come to play a justificatory role in my deliberation on account of my valuing helping, being his friend, being a compassionate person, and so on. When we cite background reason-enabling conditions we deepen the basic understanding of why the agent deliberated and acted as she did that is provided by citing her reason for acting, but we do not replace or debunk that basic understanding. Rather we make explicit the background story against which her so reasoning makes sense.[26]

Just as there can be background reason-enabling conditions, there can be background reason-*dis*abling conditions. Some role-related norms that we accept regarding what considerations to take into account in our role-based deliberation are of this form; for example, in my role as distributor of public famine relief, I bracket entirely the consideration "she's my daughter" and insulate it from playing the justificatory role it normally plays in my deliberation. Some background conditions are explicitly context linked, as are role related norms;

others are open-ended and enable or disable in any deliberative context, as does compassion.

Agents can take the same consideration to be reason-giving because of different background reason-enabling conditions. To build on an example of Michael Bratman, the members of a college admissions committee might agree to treat the consideration that a prospective student has a family tie to the institution as reason-giving, but one does so because it brings fundraising benefits that she finds important, another because she thinks it honors an implicit promise to alumni, and a third simply because she recognizes the importance of being able to co-deliberate on the basis of a shared understanding of relevant reasons.[27]

With the notions of "taking a consideration to be reason-giving" and "background reason-enabling conditions" clarified, let's return to the problem of how to characterize the motivational structure of the trustworthy. Perhaps the most influential account of trustworthiness is that implicit in Annette Baier's seminal discussion of trust, according to which trust is entrusting on the basis of belief in the goodwill of the one-trusted. It is not trust if we rely on the other's "dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all".[28] Baier does not define what goodwill is; but clearly if it is identified with friendly feelings then the account would be far too restrictive. When Jones developed a goodwill-based account that differed from Baier's in also

emphasizing affect and expectation, she too failed to define goodwill, noting only that:

> There are a number of reasons why we might think that a person will have and display goodwill in the domain of our interaction with her. Perhaps she harbours friendly feelings towards us; in that case, the goodwill is grounded on personal liking. Or perhaps she is generally benevolent, or honest, or conscientious, and so on.[29]

If we weaken the notion of "goodwill" so that it encompasses benevolence, honesty, conscientiousness, integrity and the like, we turn it into a meaningless catch all that merely reports the presence of some positive motive, and one that may or may not even be directed towards the truster. Perhaps, as Baier suggests, that which is ruled out as a motive for the trustworthy is simply anything "compatible with ill will". But then we have to define what we mean by "ill will," for on some readings it is compatible with, though never conducive to, being trustworthy with respect to a person in a domain: on a particularly vexatious morning I find myself snarling misanthropically at the whole world, yet I can still come through for some of those who are counting on me, even if not with a smile.

The notion of a background-reason enabling condition shows how we might make sense of this otherwise grab-bag list of motives. If I have robust goodwill towards someone, of the kind found in friendship or good collegial

relations, I will take the fact that they are counting on me to be a reason in my deliberation and in my action. Indeed, my so doing is partly constitutive of what it is to be a good friend or colleague. If I do not take the fact that they are counting on me to be reason-giving I am neither good friend, nor good colleague. In certain other roles, such as that of physician or teacher, my conscientious will explain why I am actively responsive to the ways my patients or students count on me. Again, being responsive in this way, within the relevant domain, is partly constitutive of being a conscientious teacher or doctor. Goodwill and conscientiousness are *different* potential background reason-enabling conditions for the *same* consideration, "he's counting on me."

Things are otherwise with fear. If I'm afraid of you, then, on pain of punishment, I must keep careful track of your expectations and strive to meet them. However, this tracking does not amount to taking the fact that you are counting on me to be a reason. The fearful agent's deliberation is tracking the prospect of retaliation and what she is being on counted on to do is simply serving as a marker for this consideration. Recall that if a condition C is reason-enabling for a consideration R, then when C is present the agent will tend to track R and to assign it a reason-giving role in her deliberation. If we vary the example so that there is no longer any prospect of retaliation for letting you down in the offing on this occasion, and I believe this, we find I would not be responsive to the fact of your dependency, notwithstanding the alleged reason-enabling condition that I'm

afraid of you still obtaining.[30] Neither fear, anger, hatred, or indifference (which seems positively *dis*abling) can function as background reason-enabling conditions for the consideration, "he's counting on me."

The key to understanding trustworthiness lies in the reason structure of the trustworthy: they take the fact that they are being counted on as reason-giving. Behind that reason structure can lay different reason-enabling conditions, including differences in motivation, but the story of trustworthiness itself is simple. The simple story not only explains how some motives get to be possible candidates for the motives of the trustworthy, while others are not, it also gives us an error theory to explain the pull of the thought that the trustworthy must have goodwill, or at the very least lack ill will. There is *a* minimal sense in which the trustworthy can indeed be said to have goodwill towards the truster: just in virtue of being responsive to the fact of someone's dependency, we *thereby* show them a measure of goodwill. The mistake is in thinking that this goodwill is something distinct from the responsiveness itself.

The notion of background reason-enabling conditions also lets us adjudicate more controversial proposals regarding the motivational structure of the trustworthy, such as those offered by Russell Hardin and Phillip Pettit, in which a motive that might be thought to be negative and perhaps trustworthiness undermining comes to be tamed and rendered social by a kind of "cunning".[31] I focus on Hardin. According to Hardin, to be trustworthy is to have an interest in

taking the interests of the truster into account, typically because of a desire to maintain that relationship. The trustworthy thus come to encapsulate the truster's interest in their own and so come to be oriented towards the truster in their deliberation and motivation. Or rather, we should say, though Hardin only sometimes does, that the trustworthy have an interest in acting on just that subset of the truster's interests that they are being counting on to advance, for trust and trustworthiness are always tacitly limited in domain. The key to maintaining an on-going relationship is to meet the expectations that the other party has for that relationship. Ignoring interests outside that area is less likely to jeopardise the relationship than is busy-body meddling in interests you were not charged with advancing. The encapsulation of interests needed here is partial, not complete, and it focuses on those interests that are also the target of expectations: "You can more confidently trust me if you know that my own interest will induce me to live up to your expectations. Your trust is your expectation that my interest encapsulate yours".[32] And – working backwards – my trustworthiness is my capacity to recognize that my interests are dependent on responding to your (success critical) expectations.[33] In other words, my trustworthiness is my being actively and positively responsive to the fact of your dependency, as the conceptual role argument requires, but mediated by the motive of self-interest, functioning in a background role.

The problem is self-interest is unable to ground genuine responsiveness to dependency. The self-interested agent's deliberation may, for a time, track the other person's expectations, but, as in the fear case, this tracking does not amount to taking the fact she is being counted on as a reason. The truster's expectations, if met, are merely a marker for that which will promote self-interest and it is this later that is playing a justificatory role in deliberation. We can see this by looking at what happens when self-interest and responsiveness come apart. Hardin's own example, drawn from *The Brother's Karamazov*, shows the self-interested do not have the right reason-structure to be trustworthy. The merchant Trifonov and an army officer have, while it lasts, a mutually profitable relationship in which they use army funds for personal gain.[34] When the officer's posting ends, Trifonov refuses to return the borrowed money and disavows the existence of their "arrangement." The encapsulated interest story is not a story of how someone else's interests become our own, and so how we come to be responsive to their dependency – for we are not, though for a time, our interests lead us to act as if we were.

3.3 Keeping it simple

So far, I have not been very precise in my formulation of the simple view, sometimes omitting specific mention of the domain, or the truster; and nowhere discussing the force that the consideration that someone is counting on you has in

the deliberation of the trustworthy. It is time to be more precise. Let's begin with a canonical statement of what I'm going to call "basic trustworthiness". Like trust, basic trustworthiness has three-place structure:

Basic trustworthiness – B is trustworthy with respect to A in domain of interaction D, if and only if she is **competent** with respect to that domain, and she **would** take the fact that A is counting on her, **were** A to do so in this domain, to be a be a **compelling** reason for action.[35]

The formulation needs unpacking. A compelling reason is not an overriding one, but it is not easily outweighed. The trustworthy (with respect to A, in D) who are called on to act on their trustworthiness, either deliver or have some excusing explanation for why they did not. This explanation could reveal that something untoward happened which prevented their competence from bringing success without casting doubt on its existence. Or it could be that abnormal circumstances threw up some yet more compelling reason that prevented them from acting to fulfil the truster's expectations. There is a necessary vagueness about what it is to take a reason to be compelling and there can be disagreement over whether an agent has in fact done this. Someone might be unfairly judged untrustworthy when they are not, and untrustworthiness can be disguised behind claims that other reasons are more pressing. Assessing trustworthiness can thus often be controversial – but this is what we should expect, rather than a problem for the account.

Though the trustworthy (with respect to A in D) take the fact that A is counting on them to be a compelling reason, they may or may not explicitly deliberate about what to do and if they do, they may express that reason in various ways. Much trustworthy behaviour becomes part of our everyday routine and we need to reflect on the fact that others are counting on us only when some temptation threatens to disrupt habit.[36] Nor need "A is counting on me" figure in so many words on those occasions when I do deliberate: I mean it as a schematic summary of the various ways in which in we might refer to the fact of the other's dependency. We might express it to ourselves or to others in terms of "following through," "expecting my help," "letting them down," "being there;" even, with enough of the right background, "it's what I always do."

Trustworthiness is dispositional. I can be trustworthy with respect to a person and a domain and yet never be called on to display my trustworthiness. Trustworthiness is expressed in action when activated by being counted on.[37] To be trustworthy with respect to A in D thus requires that B be capable of recognizing that A is counting on her and, roughly, what they are counting on her for. B is not trustworthy if she acts when she *thinks* A is counting on her when A is doing no such thing. She needs to have a disposition that is keyed to A's counting on her and so activated when that happens. Perhaps B might go wrong here sometimes without losing her claim to trustworthiness, but there is both an excess and a deficiency that undermines trustworthiness. One can be either overly

prone to thinking others are counting on you, or insufficiently prone. The former will tend to be untrustworthy because officious and meddlesome. The later will routinely drop the ball: "What, you were expecting me to catch it? Oops! Sorry." As well as being mistaken about whether someone is counting on you, you can be mistaken about *what* they are counting on you for. It is rarely as clear as doing a specific action, though it can be. Often it is some rather vaguely specified broader project that they are counting on your helping them advance, or some good they hope you will care for. It takes attunement to others to grasp these things; typically, though not invariably, it takes a kind of social ability that extends beyond the capacity to respond to a particular agent. This shows the role of background social knowledge in being trustworthy and explains why it can sometimes be hard to be trustworthy for someone from a radically different cultural background. One would respond if only one knew when and how.

Once unpacked, the definition of basic trustworthiness suggests strategies for promoting it. There are two different kinds of strategies that can be pursued. First, we can increase the prevalence of interactions in which a background reason enabling condition for responsiveness to the fact of dependency will be present. For example, we can design institutions that foster conscientiousness on the part of those in institutional roles. Second, we can reduce the field of competing considerations so that responsiveness to dependency will more often carry the day. Ordinary flawed human beings have basic trustworthiness with respect to

many domains in their interaction with other human beings. We are 'almost trustworthy' with respect to a great many more. It is part of our common humanity, grounded in our capacity for sympathy, that we are susceptible to being responsive to the dependency of others. The problem is not getting us to recognize dependency as a reason, but rather getting us to give it enough weight so that it can become a compelling reason. We are poised, as it were, to be trustworthy if only doing so were compatible with other things we also care about. Any institutional or interpersonal strategy that reduces conflict of interest will, all by itself, enhance the trustworthiness of the almost trustworthy, and it may not take much to tip them over the line into trustworthiness proper.

When we talk about cultivating trustworthiness, it is often more than basic trustworthiness that we have in mind. We want rich trustworthiness that correctly signals its presence. (And, of course, we want trustworthiness in service of good ends to be in plentiful supply, while trustworthiness in service of bad ones is all used up: more on this in the next section). Someone might have basic trustworthiness with respect to A in a domain, yet never be called on to display it. This may be through no fault of their own: potential trusters might have been scared off trusting where they legitimately might by stereotype and prejudice. If that is so, then the failure of their trustworthiness to receive proper uptake and recognition in trust is itself a form of disrespect. Sometimes, though, the failure of trustworthiness to receive uptake is a fault of the trustworthy – while they have

basic trustworthiness we want something more from them. We want them to reveal their trustworthiness, and not through words, for as Baier reminds us:

> "Trust me!" is for most of us an invitation which we cannot accept at will – either we do already trust the one who says it, in which case it serves at best as reassurance, or it is properly responded to with, "Why should and how can I, until I have cause to?"[38]

We want them to signal their trustworthiness in a domain by "walking the walk," by showing us that they are competent and can be counted on by actually *doing* something that anticipates the ways in which we would want to be able to count on them, if only we knew we could. Unsignalled or unreliably signalled trustworthiness is no use to us.[39]

Trustworthiness that reliably signals its presence (rich trustworthiness) requires capacities significantly more sophisticated than those required for basic trustworthiness (which are themselves not trivial). Correctly signalling my trustworthiness (to a person regarding a domain) requires grasping what the other will count as a signal. Signalling rests on a set of highly complex socially-mediated background understandings. These provide a framework in which, like it or not, we are always already signalling what we can be counted on for. Individual competence in signalling requires understanding what is being signalled to whom through these socially-mediated "standing channels" and

knowing how and when to override that signal in order to communicate that I can be counted on for more than, or less than, might be expected.

Rich trustworthiness requires not only competence in a domain, but also competence in assessing my own competence, so that I neither signal competences I do not have, nor "hide my light under a bushel." I need to engage in on-going reflective self-monitoring of my own competences so that I know them and their limits.

Though harder to cultivate than basic trustworthiness, there is much that we can do to scaffold our own and other people's ability to be richly trustworthy. Our capacity to monitor our own competence can be scaffolded both interpersonally and institutionally. Certification boards and watchdogs can contribute to securing competence and accurate self-perception of competence. When working properly, they signal the role-based competences of those they certify. Friends hold up a mirror in which we can more accurately view our own strengths and limitations, so that our self-monitoring need not be conducted alone.

Rich trustworthiness requires the coordination of a sophisticated set of competences: in domains, in self-assessment, in signalling, and in the practical wisdom required to be alive to the expectations of others and appropriate ways in which they might be met. The concept of trustworthiness has an indispensable normative role because it helps us assemble and sustain the relevant competences. Without it playing an explicit role in our moral education it would be impossible

for us to develop this complex suite of capacities. Sympathy gives us the capacity to be responsive to the fact of other people's dependency, but it is through our early interactions with others that we become trustworthy and through our on-going interaction with them that we are sustained in our trustworthiness as our trustworthiness receives uptake in trust. We respond to "hopeful" trust in which the other holds out an image of ourselves as competent and responsive and we enact the selves they see us as being.[40] It is in part because we and others have the concept trustworthy that we become trustworthy; trustworthiness is thus built through the looping mechanism characteristic of human kinds.

The simple account of trustworthiness underwrites a guarded optimism about the prospects for trustworthiness in contemporary life. Several features of modern urban living might be thought to support pessimism about trustworthiness and hence about the wisdom of trust. Better, goes a common view, to economize on trust, since trustworthiness can be predicted to be in short supply in complex, anonymous, pluralistic societies. In face-to-face societies, where interactions are largely between people one knows, or people in known relations to known others, shunning and shaming provide strong incentive to follow through on conventions and expectations. Rich overlapping social networks undergird the goodwill characteristic of communal or kinship relations between many of the people with whom one must interact. Relationships are typically long and other people's interests can come to be deeply embedded in one's own in virtue of this. Perhaps

most significantly, members of smaller less diverse societies are more likely to share fundamental evaluative outlooks. None of these conditions hold in most developed urban societies: we know there is significant divergence in values in pluralistic societies; people are more mobile, meaning relationships are shorter, reputational effects reduced. Thus if trustworthiness requires shared values, encapsulated interests, or goodwill, the prospects for it being widespread in contemporary urban societies look bleak. Better then, to come up with cunningly designed institutions so that we can economise on trust before our cash reserves of trustworthiness run out.

The account of trustworthiness defended here is more optimistic: one needs neither goodwill (except in the minimal sense associated with responsiveness itself), nor on-going relationships, nor even shared values. Trustworthiness cannot be elicited in the service of ends that you actively *disvalue*, but you need not share common values to be capable of responding to the fact of another's dependency. Sometimes, the fact that they are counting on you can, all by itself, be enough.

4. Is rich trustworthiness a virtue?

Trustworthiness does not rate a mention on classical lists of the virtues. Loyalty, trustworthiness's close relative and a ground on which trustworthiness can be demanded is on many classical lists, but it is nowadays treated with suspicion as a

virtue that makes sense only in stratified societies where it functions to keep the serf in his place and the wife in hers.

The standard case against trustworthiness as a virtue or as morally required is fourfold:

1. Trustworthiness can be in the service of bad ends as well as good ones. Evil thrives when evildoers work together. Thus,
2. One can be required to respond to trust, extended in service of evil ends, with "trust busting".[41]
3. It is not always wrong to actively elicit trust and then "bust" it with treachery.
4. There need be no fault in refusing to respond to unsolicited trust with trustworthiness, for sometimes trust can itself be an imposition.[42]

Lawyers for the defence can, however, point to recognized virtues that show similar features, arguing that, in the company of such noble partners, there can surely be no crime. Courage can be used in the service of bad ends as well as good ones. Or if you prefer to say instead that courage in the service of unjust ends is no true courage, and that to have true courage one must also have the virtue of justice, then trustworthiness must, in fairness, be allowed the same defence. The honest can sometimes be required to lie. A spy infiltrating the command of a genocidal enemy will need to dissimulate and lie, yet might still claim the virtue of honesty. They might actively seek a reputation for honesty among the enemy in order that their lies might be more readily believed. The

prosecution's final argument is a little harder to rebut, for surely, say, within the limits of justice and capacity we are required to respond to need with benevolence. But lawyers for the defence remind us that it is rich trustworthiness that is a candidate for a virtue and not basic trustworthiness. The richly trustworthy will neither give false signals regarding who can trust them and for what (spy-cases aside), nor will they merely turn their backs on unsolicited trust. They will indicate that it is misplaced and invite it to be withdrawn.[43]

Trustworthiness leaves the dock.

The prosecution needs a stronger team, for the problem with trustworthiness lies deeper than has so far been tested. Recall that a virtue is an excellence of character that comprises a stable suite of dispositions to action, feeling, perception, and to recognizing the reasons characteristic of the virtue in practical deliberation. Honesty, for example, requires recognizing the importance of taking "it is the truth" as a reason but, unlike the rude or garrulous, the honest combine that recognition with the perceptual discernment to negotiate social conventions regarding truth-speaking.

Could trustworthiness have the right kind of dispositional structure and stability to be a virtue? Not so long as we are thinking of it in three-place terms. For human beings, with our patchwork of competences, there can be no generalized disposition to trustworthiness: even if I am generous in my responsiveness to the dependency of many different others, the limits of my

competence will mark the limits of my trustworthiness. Things are otherwise with honesty or benevolence, for the competences they require are themselves part of the virtue. If you do not have the social competence to discern what counts as honesty, what rudeness, then you have yet to acquire the virtue and you are, to that extent, ethically deficient. You are not ethically deficient if you lack the competence to be trustworthy with respect to many domains.

The argument is not over, yet. Perhaps trustworthiness-the-virtue lacks three-place structure. Rich trustworthiness itself involves the ability to monitor and signal one's competences, and if I've done that, then we could argue that I've been trustworthy, even though I cannot be relied on to act as someone might expect me to in a particular domain.[44] The domain problem raises suspicion, but we have to look elsewhere to prosecute the case.

We find our clue by noticing a difference between trustworthiness and other recognized virtues, which leads us to revisit our earlier concession that trustworthiness and honesty behave the same. Consider the following propositions:

1. We can be required by justice to tell a lie.
2. We can be required by justice to let down someone who is counting on us.

Both make sense and, so far, the parallel seems to hold. Now consider:

3. We can be required by honesty to tell a lie.

4. We can be required by trustworthiness to let down someone who is counting on us.

(4) makes sense, while (3) simply does not. Return to our spy: to be trustworthy with respect to the liberation army, they must be untrustworthy with respect to the genocidal enemy. The conflict is internal to the alleged virtue itself. The state itself is potentially inwardly riven, because it takes as its signature reason a consideration that has inherent possibilities for conflict with the self-same consideration. "He's counting on me" – but all sorts of people can count on us for all sorts of incompatible things. Nor need this incompatibility be merely contingent, analogously with the contingent incompatibility of helping two people at the same time with resources sufficient only for one. The spy is being counted on for two contradictory things: help advancing the genocidal cause and help stopping it in its tracks.

At this point the defence team reminds us of loyalty. Given loyalty's hold on current lists of the virtues is contested, trustworthiness is no longer keeping irreproachable company.[45] But things are even worse. If loyalty is to keep the title "virtue" it its grip, it cannot be thought of as blind obedience to a community or cause with which one simply finds one's life and sense of identity bound up. It needs to be thought of as keeping faith with a *commitment*, perhaps as embodied in community.[46] Our commitments can conflict contingently and we must chose between them, just as there can be conflicting demands on our benevolence.

However, once commitments conflict intrinsically, we are required to withdraw from one or the other. Once we withdraw from a commitment, loyalty – understood as keeping faith with a commitment, not blind obedience – loses its grip. Our new loyalties require us to abandon the old, but this is only sloppily expressed by saying, "loyalty requires us to be disloyal," for once a commitment is abandoned we cannot be disloyal to it. Things are otherwise with trustworthiness precisely because the source of its characteristic reason lies in other people and their expectations of us.

If these reflections are along the right lines, trustworthiness is not a virtue: it is no accident that it has been left off all the classical lists. It names something important, something we care about but part of the reason we care about it is because it resists moralization. As finite dependent social beings who are often less than fully virtuous, we want there to be people who will respond to what we are counting on them doing, even when what we are counting on them doing is not fully virtuous.

---

[1] There are two notable exceptions. Nancy Potter begins from the trustworthiness end in *How Can I be Trusted? A Virtue Theory of Trustworthiness* (Lanham, Maryland: Rowman and Littlefield, 2002) and Russell Hardin proceeds by investigating trust and trustworthiness in tandem, see *Trust and Trustworthiness*

(New York, NY: Russell Sage Foundation, 2002). More about these theorists later.

[2] For quite different affect-based accounts, see Richard Holton "Deciding to Trust, Coming to Believe," *Australasian Journal of Philosophy* 72-1 (1994): 63-76, Karen Jones, "Trust as an Affective Attitude," *Ethics* 107 (1996): 4-25 and Lawrence Becker, "Trust as Noncognitive Security About Motives," *Ethics* 107: 4-25. The best known belief account is Hardin, *Trust and Trustworthiness;* while John Dunn, "Trust and Political Agency" in *Trust: Making and Breaking Cooperative Relations*, ed. Diego Gambetta (Oxford: Blackwell, 1988): 73-93, and Virginia Held, "On the Meaning of Trust," *Ethics* 78(1966): 156-59 emphasize action-like features of trust, and Annette Baier, "Trust and Its Vulnerabilities" in her *Moral Prejudices* (Cambridge MA: Harvard University Press, 2002) 130-151 offers a combined account.

[3] See Annette Baier, "Trust and Antitrust," *Ethics* 96(1986): 231-260 and Jones, "Trust as an Affective Attitude."

[4] Hardin *Trust and Trustworthiness*, 3.

[5] Hardin *Trust and Trustworthiness*, 13-14; see also his discussion of theorists that allegedly confuse trust and acting as if you trust, 58-60. For a different argument to the same conclusion that draws on general considerations about what can serve

as a reason for what kinds of attitudes, see Pamela Hieronymi, "The Reasons of Trust," *The Australasian Journal of Philosophy* 86-2(2008): 213-236.

[6] See Justin D'Arms and Daniel Jacobson, "The Moralistic Fallacy: On the 'Appropriateness' of Emotions," *Philosophy and Phenomenological Research* 61-1(2000): 65-90.

[7] See D'Arms and Jacobson, "The Moralistic Fallacy" and their "The Significance of Recalcitrant Emotion (or, Anti-Quasijudgmentalism)," *Philosophy* (supp) 52 (2003): 127-145.

[8] Natural and social kinds have such explanatory roles. See Richard Boyd, "Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61(1991): 127-148.

[9] See Partha Dasgupta, "Trust as a Commodity." In *Trust* Gambetta ed. 49-72 and Francis Fukuyama, *Trust: The Social Virtues and the Creation of Prosperity* (New York, NY: The Free Press, 1995). For a critique of these and related explanatory claims, see Hardin, *Trust and Trustworthiness*, 81-84.

[10] Karen Jones, "Trust: Philosophical Aspects," in Neil Smelser and Paul Bates (general eds.), *International Encyclopedia of the Social and Behavioral Sciences*; (Amsterdam: Elsevier Science, 2001), 15917-922.

[11] For reactive attitude accounts, see Holton "Deciding to Trust" and Margaret Walker, *Moral Repair: Reconstructing Moral Relations after Wrongdoing*

(Cambridge: Cambridge University Press, 2006) Chapter 3. In "Trust and Anti-trust," Baier offers a goodwill account, as does Jones in "Trust as an Affective Attitude." For an account focusing on integrity, see Carolyn McLeod, *Self-Trust and Reproductive Autonomy* (Cambridge MA: MIT Press, 2002).

[12] See Lars Hertzberg, "On the Attitude of Trust." *Inquiry* 31(1988): 307-322, H.J.N Horsburgh, "The Ethics of Trust," *Philosophical Quarterly* 10(1960): 343-354 and Olli Lagerspetz, *Trust: The Tacit Demand* (Dordrecht: Kluwer, 1998).

[13] Baier, "Trust and Anti-trust," 232.

[14] Ian Hacking, "On Boyd," *Philosophical Studies* 61 (1991): 149-154, 154. Boyd himself acknowledges these roles, so the dispute here is more apparent than real.

[15] Ian Hacking, *Rewriting the Soul: Multiple Personality and the Sciences of Memory* (Princeton NJ: Princeton University Press, 1995).

[16] For an insightful discussion of the constructive role of gender norms, with references to the broader literature, see Sally Haslanger, "On Being Objective and Being Objectified," in *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, ed., Louise Antony and Charlotte Witt (Boulder, CO: Westview Press, 1993), 85-125.

[17] For arguments with this structure see Sally Haslanger, "Gender, Race: (What) Are They? (What) Do We Want Them To Be?" *Noûs* 34:1 (2000): 31-55 and "What Knowledge Is and What It Ought To Be: Feminist Values and Normative Epistemology," in *Philosophical Perspectives* 13 (1999): 459-480. See also,

Justin D'Arms, "Two Arguments for Sentimentalism," *Philosophical Issues* 15 (2005): 1-21.

[18] See Philip Pettit *The Common Mind: An Essay on Psychology, Society and Politics* (New York: Oxford University Press, 1993) Chapter 2 for a discussion of the capacities required.

[19] This is the story behind trust-responsiveness. A full discussion of trust-responsiveness is beyond the scope of this article, but see notes 33 and 39.

[20] I do not mean to claim that trust and trustworthiness are the whole, or even the chief part, of the story of human sociality. They are a *significant* part of it. Perhaps we use also recruit our sociality to solve the problem of our finitude through sharing intentions; see Michael Bratman, "The Dynamics of Sociality," *Midwest Studies in Philosophy,* XXX (2006): 1-15 and "Shared Intention," *Ethics* 104 (1993): 97-113. I suspect, however, that trustworthiness often stands in the wings to keep shared intention working when commitment wavers, so that it is indeed central to understanding our sociality, but my argument does not hinge on this.

[21] Diego Gambetta, "Can we Trust Trust?" in his, *Trust: Making and Breaking Cooperative Relations*: 214-237, 217-18, original emphasis.

[22] There are even thinner accounts of trust, according to which trust is reliance with or without confidence. I leave these accounts to one-side, however, on the

grounds that they make trust identify too heterogeneous a class of dependencies to sustain any useful generalizations (see 1.2). It is interesting that risk assessment accounts of trust, the second thinnest, identify a category that can sustain generalizations and do explanatory work, but its partner trustworthiness concept does not. If it is reasonable to demand that both partnered concepts must earn their keep, then we can see how we get extra leverage by switching to work the problem of understanding them to the trustworthiness end.

[23] See Scanlon, *What We Owe to Each Other* (Cambridge MA: Harvard University Press, 1998), 17-20. For an argument that this is not the best way to characterize the "reason-for" relation, see Pamela Hieronymi, "The Wrong Kind of Reason," *The Journal of Philosophy* 102-9 (2005): 437–57. Instead of talking of "considerations" others prefer to speak of reasons as facts; for example, see Niko Kolodny, "Why Be Rational?" *Mind* 114 (2005): 509-563. My interest is in what it is to take something to be a reason, rather than what it is to be a reason, thus nothing in my argument hinges on the details of how we understand the "reason-for" relation. I'll work with Scanlon's simpler formulation and speak indifferently of "taking the fact that-" and "taking the consideration that-" to be reason-giving.

[24] This discussion of taking a consideration to be reason-giving is indebted to Scanlon, "Reasons and Passions', in Sarah Buss and Lee Overton eds. *Contours*

of Agency (Cambridge MA: MIT Press, 2002), 165-183. Scanlon writes: "…more frequently than I would like, I count as reasons-for action, or for other attitudes, considerations that I actually believe do not, under the circumstances, count in favor of those attitudes. Considerations can seem to me to be reasons even when I have judged that they are not." (170). And, we might add, they can seem to me to be reasons without my taking them to be so.

Is "he's counting on me" *really* a reason? Officially, I mean to remain neutral about this, as the answer depends on other commitments in your preferred theory of reasons. The conceptual role argument establishes only that we have a strong interest in there being agents who will take "he's counting on me" to be reason-giving, not that being so counted on provides a reason, let alone a non-optional reason that must be taken into account. For the record, I think that sometimes it is, and sometimes it isn't: it all depends on who is being counted on for what and by whom. But it is open to someone to say it is a reason, but sometimes its force is outweighed.

[25] See Bratman, "Dymanics" where he defines values in terms of policies to treat considerations as reason-giving, rather than evaluative judgments. For a discussion of accepting norms regarding considerations to treat as reasons, see Allan Gibbard, *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge, MA: Harvard University Press, 1990) 160-64.

[26] Consider compassion: a compassionate person will take the fact that someone needs their help to be a reason to help them. Let us suppose that, in a particular case, were they not compassionate, they would not help. Should we then infer that the reason for their helping is not "they need my help", but rather "I am compassionate"? That would be to accept the slander that the virtuous act with one eye on their own excellence of character, which, we might think, would preclude them from having any. Compassion is a background reason-enabling condition that explains why considerations like "he needs my help", "she's lonely" and so on, are taken by an agent to be reason-giving.

[27] Bratman, "Dymanics," 5.

[28] Baier, "Trust and Anti-trust," 234.

[29] Jones, "Trust as an Affective Attitude," 7.

[30] There can be mixed cases, too; e.g. responding to the expectations of a stern but revered father figure.

[31] Hardin *Trust and Trustworthiness*, and Pettit "The Cunning of Trust," *Philosophy and Public Affairs* 24 (1995): 202-25. My focus will be on Hardin, but a similar point applies to Pettit. Love of esteem only loosely anchors an agent to what people are counting on them to do, since the considerations readily pull apart. I think love of esteem is sometimes part of the story of trustworthiness, but

it works quite indirectly, explaining why we tend to like those who (seem to) like us. Amiability, activated by liking, is indeed reason-enabling for "he's counting on me." If we think our amiability is being exploited it quickly fades. For a cogent argument that esteem-seeking is a manipulative and inherently unstable motive on which to ground trust-responsiveness, see Victoria McGeer, "Trust, Hope and Empowerment," *Australasian Journal of Philosophy* 86-2 (2008): 237-254.

[32] Hardin *Trust and Trustworthiness*, 5.

[33] See *Trust and Trustworthiness,* 28, where Hardin makes it explicit that the key is responding to what "one is trusted to do."

[34] *Trust and Trustworthiness*, 1-3.

[35] I am going to work with the "domain of interaction" formulation, rather than the more popular "with respect to action Z", because I think even basic trustworthiness has a certain "breadth" and thus must extend beyond the performance of a specific action. The case for preferring the domain formulation is stronger with respect to trustworthiness than it is with respect to trust, where the action formulation has currency, but even there I think a case for it can be made and not merely on the general methodological grounds that the two concepts are paired and so should display complementary structure. For the domain formulation with respect to trust, see Jones "Trust as an Affective Attitude."

[36] Baier rules out "dependable habits" as a motive for trustworthiness in "Trust and Anti-trust," 234. But we can have habits of trustworthiness, too.

[37] Nevertheless, trustworthiness is not always a matter of trust-responsiveness. A discussion of what partner concept of trust best fits with this account of trustworthiness is beyond the scope of this article, but for the record, I think trust is not best seen as merely "counting on" (recall from section 2 we can count on things as well as people) but rather as "expecting the other to be directly and favourably moved by the thought that you are counting on them" (see Jones 1996). Thus trustworthiness can be shown where the other does not (or does not yet) trust. Take the now familiar example of Kant and his neighbours who use his regular habits to tell the time. Kant probably is trustworthy with respect to providing the time to his neighbours because, being an obliging sort of person, he would take the fact that they are counting on him to be reason-giving. But his regular habits are not currently evincing that trustworthiness. Suppose he came to know that they depended on him in this way, then the regularity of his habits could come to express his trustworthiness as that consideration received uptake in his practical deliberation. Suppose further that it became common knowledge among his neighbours that Kant was aware of their habitual reliance, then they would come to count on his responding to their counting on him. This iterated

dependency is at least *a* source of the normativity of conventions, but defending that thought is a job for another occasion.

[38] "Trust and Anti-trust," 244.

[39] Compare Potter who identifies the core dispositions of those who posses trustworthiness as a virtue to be "They give signs and assurances of their trustworthiness" and "They take their epistemic responsibilities seriously," *How Can I be Trusted?*, 174-75.

[40] See Victoria McGeer, "Trust, Hope and Empowerment."

[41] Baier, "Trust and Anti-trust," 232.

[42] Jones, "Trust as an Affective Attitude," 9.

[43] Potter, *How Can I be Trusted?*, 26-27.

[44] In *How Can I be Trusted?* Chapter 1, Potter makes a similar point about "full trustworthiness", which she claims does not have three-place structure. It requires, however, a commitment to a specific set of liberatory egalitarian values, rather than responsiveness to other people's counting on you. I think her description of trustworthiness as a virtue is in fact a description of what it would take to be trustworthy with respect those who shared similar values.

[45] For a helpful discussion see Keller, *The Limits of Loyalty* (Cambridge University Press, 2007), Chapter 2.

[46] The source of the idea that loyalty is commitment to a cause as embodied in a community is Josiah Royce, *The Philosophy of Loyalty* (New York: MacMillan, 1908). Royce emphasizes the importance of judging the cause to be worthwhile and choosing it freely rather than being merely uncritically committed to those communities into which one is born. See Keller, *The Limits,* for a contemporary discussion that sees an important moral difference between chosen and found loyalties.