Trustworthiness
Author(s): Karen Jones
Source: *Ethics*, Vol. 123, No. 1 (October 2012), pp. 61-85
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/10.1086/667838
Accessed: 17/05/2013 15:43

Trustworthiness*

Karen Jones

I present and defend an account of three-place trustworthiness according to which 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 compelling reason for acting as counted on. This is not the whole story of trustworthiness, however, for we want those we can count on to identify themselves so that we can place our trust wisely.

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 comparative silence, we can discern 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 backward like writing viewed in a mirror.

In this article, I reverse the focus and approach the problem of understanding trust and trustworthiness from the trustworthiness end, leaving trust to be glimpsed reflected through the mirror of trustworthiness. Trustworthiness is interesting and worthy of investigation in its own right. But part of the point of switching to approach the problem of understanding trust and trustworthiness from the trustworthiness end is that the normative role of these concepts comes more clearly into view when approached from this direction. The resulting account of trustworthiness

* This essay has gathered many debts on its long route to publication, most recently to the editors of, and two referees for, this journal. Older debts are to audiences at the University of Toronto, Universite de Montreal, Dubrovnik, Aarhus University, and a University of Helsinki sponsored workshop in Rome, which first got me thinking about trustworthiness.

[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, MD: Rowman & Littlefield, 2002), and Russell Hardin proceeds by investigating trust and trustworthiness in tandem (see *Trust and Trustworthiness* [New York: Russell Sage, 2002]). More about these theorists later.

Ethics 123 (October 2012): 61–85

will have implications for how we should understand trust, but these implications cannot be explored here, where my focus remains squarely on trustworthiness.

It is common practice for philosophers to use intuitions about examples to defend their preferred accounts of trust and to critique rival accounts. I do not take that approach in my investigation of trustworthiness. Instead, I begin by asking why we have the paired concepts trust and trustworthiness, arguing that trustworthiness and trust are not reducible to reliability and reliance because they identify, in order to promote, a distinctive way that our cognitive sophistication makes it possible for us to respond to the fact of interpersonal dependency. I argue that this role supports an account of trustworthiness—in a domain, with respect to an agent—as competence together with direct responsiveness to the fact that the other is counting on you. Trustworthiness, so understood, has three-place structure. Three-place trustworthiness is not, however, the whole story of trustworthiness. As finite agents, we want more from others than responsiveness to dependency in a domain: we want those who can be trusted to identify themselves so that we can place our trust wisely. I label this further dimension "rich trustworthiness." Rich trustworthiness names something less than a virtue, but it identifies a trait that 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.

I. A 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 nonhuman 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.[2]

[2] 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. But as will become apparent in Sec. IV, I think that the end result is not especially revisionary. Adopting the strategy of conceptual role argument is able to finesse the question of whether we, the folk, have just one pair of trust/trustworthiness concepts or many. I think that ordinary usage is quite messy and probably does group together a number of somewhat distinct, more or less closely related, trust/trustworthiness concepts, but this does not count against the merits of an account that is able to fulfill a clearly defined conceptual need. For conceptual role arguments, see Sally Haslanger, "Gender, Race: (What) Are They? (What) Do We Want Them to Be?" *Noûs* 34 (2000): 31–55. See also Justin D'Arms, "Two Arguments for Sentimentalism," *Philosophical Issues* 15 (2005): 1–21.

The conceptual role argument provides a test against which competing accounts of trustworthiness (and of trust) can be checked for adequacy.

Trust and trustworthiness have distinctive conceptual work to do because of three fundamental facts of human existence: we are social, finite, and reflective 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. Some tasks cannot be accomplished solo because they are so large they require many hands; other tasks require a variety of skills and competences that cannot feasibly all be had in adequate measure by any one agent. Sometimes the problem is simply one of time: we cannot be everywhere or do everything. The fact of our sociality means that we have available to us a potential of pool of other agents whose competencies and effort can be recruited for our ends or for shared projects. Sociality is not only a source of risk, it is also a source of power. Our reflectiveness enables a distinctive way in which the risks of sociality can be reduced and its power harnessed.

As reflective beings, we have the potential to stand in a distinctive kind of relationship to one another. 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 from, and harness the power of, natural forces by taking measures to control them: canals and embankments reduce the risk of flooding; irrigation channels take water from where we don't need it to where we do. However, the measures we can take to control and harness 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 behavior nor the capacity to anticipate our behavior. Only agents, whether human or animal, have the ability to control and to anticipate.

Animal agents have the capacity to modify their behavior in the light of their anticipation of the behavior of others and so can pursue behavioral strategies that embed assumptions about what other creatures will do. Dogs do this when they respond to other dogs signaling their intention to attack, to submit, or to play. In this sense, their behavior can depend on their understanding of the behavior 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.[3] 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 and goals an expectation that, if false, means you risk being left worse off than you otherwise would have been. The success of your plan or the achievement of your goal depends, nontrivially, on what you are counting on coming to pass. 'To count on' is thus to do something more than merely to expect: we have all sorts of expectations about the behavior of things and people, but only some of them come to be embedded in our plans and goals in ways that affect their success. These are the things we count on. We can count on others without being aware that we are doing so. Though aware of our plan, we sometimes become aware of the things we were counting on for its execution only when their absence wrecks the plan. We need be explicitly aware neither of our goals nor of the presuppositions for their fulfillment. We count on others for all sorts of things: some of them are practical, such as their help looking after something we care about. Others are theoretical—projects of inquiry which may or may not have significant practical implications. In many cases, it is our counting that creates a dependency: the success of our plan is now dependent on, or hostage to, the behavior of the thing or person that we count on. Dependencies can be unwitting: had we recognized the assumption embedded in our plans and thought it unlikely to be fulfilled, we would not have gone ahead. Dependencies can also be inescapable, 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.[4]

[3] See Philip Pettit, *The Common Mind: An Essay on Psychology, Society and Politics* (New York: Oxford University Press, 1993), 54–76, for a discussion of the capacities required.

[4] This is the story behind trust responsiveness. A discussion of trust responsiveness is beyond the scope of this article, but see note 16.

Knowing that others can themselves recognize and respond to our dependency means we can actively seek to recruit their agency and their competencies to enhance the effectiveness of our own. Of course we do not always do this. Sometimes we treat other agents much like natural forces whose behavior 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 behavior 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 often enough 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 nonreflective 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 interhuman 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 both to reduce the risk of dependency and to enable us to harness its power. We are thus interested in labeling that distinctive mode (whatever exactly it is) of engaging with the other's dependency in order to promote it.

II. TRUSTWORTHINESS AS ACTIVE ENGAGEMENT WITH DEPENDENCY

Suppose we take 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. The simplest account of trustworthiness supported by the conceptual role argument thus keeps this neutrality.

As a first approximation, someone who is trustworthy, with respect to a person in a domain, takes the fact that they are being counted on to be a reason for acting as counted on in their motivationally efficacious practical deliberation. There may be a deeper story about the agent's background motivational states, values, or commitments that explains why they are responsive to dependency in this way, but there does not have to be, and it need not always be the same story. The account thus sidesteps significant controversy regarding the motivational structure that trust tacitly imputes to the trustworthy by getting us to focus instead on the characteristic reason the trustworthy are responsive to. The account is not only simple, it also: (*a*) explains what unifies the various motivations, from goodwill to conscientiousness to integrity, that have been put forward as possible motives for the trustworthy; (*b*) explains what unifies the various states, such as fear and anger, that are typically thought not to be conducive to trustworthiness and what unifies states such as indifference and hatred commonly taken to be trustworthiness incompatible; (*c*) gives us an error theory that explains the grain of truth but also the mistake in the common association between trustworthiness and goodwill; and (*d*) lets us adjudicate more controversial proposals regarding the motivational structure of the trustworthy, such as Hardin's encapsulated interest account that makes trustworthiness a matter of enlightened self-interest.[5] The account is to be preferred for these reasons, or so I will argue.

Perhaps the most influential account of trustworthiness is that implicit in Annette Baier's generative 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."[6] 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 I developed a goodwill-based account that differed from Baier's in also emphasizing affect and expectation, I 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 harbors 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."[7]

If we weaken the notion of "goodwill" so that it encompasses benevolence, honesty, conscientiousness, integrity, and the like, we turn it into a meaningless catchall that merely reports the presence of some positive motive, and one that may or may not even be directed toward 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.

There is, though, a kind of unity to this otherwise grab bag list of motives. If I have robust goodwill toward 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 to act as I am being counted on in my motivationally efficacious deliberation. Indeed, my so doing is partly constitutive of what it is to be a good friend or colleague. If I am not responsive to the ways in which I am being counted on, I am neither good friend nor good colleague. In certain other roles, such as that of physician or teacher, my conscientiousness will explain why I am actively responsive to the ways my patients or students count on me. I am responsive to them qua my patients, or my students. Again, being responsive in

[5] Hardin, *Trust and Trustworthiness*.

[6] Annette Baier, "Trust and Anti-trust," *Ethics* 96 (1986): 231–60, 234.

[7] Karen Jones, "Trust as an Affective Attitude," *Ethics* 107 (1996): 4–25, 7.

this way, within the relevant domain, is partly constitutive of being a conscientious teacher or doctor. In many role-related contexts, the responsiveness is toward a group of people (e.g., the senator to her or his constituents), or toward individuals insofar as they are members of a group. In friendship and other intimate relationships, the responsiveness is targeted at a particular individual in virtue of properties that may be unique to her or him. Though otherwise quite different, goodwill and conscientiousness are alike in one respect: it is constitutive of having goodwill or conscientiousness that, in certain contexts, the fact that someone is counting on you can, all by itself and without further incentive, activate responsiveness.

Things are otherwise with fear. It is not partly constitutive of being afraid of someone that, in certain contexts, the fact that they are counting on you, all by itself and without further incentive, activates responsiveness. Grant that, if I am afraid of you, I might need to keep careful track of your expectations and strive to meet them, but I will do this only where there is the added incentive of avoiding your retaliation against me should I let you down. Treacherous actions that do not risk retaliation remain on my deliberative agenda, or if they do not, that is because of considerations other than my fear. In contrast, in the relevant domains, treacherous actions no longer remain on the deliberative agenda of the conscientious or the goodwilled.

As a possible motivational foundation for trustworthiness, indifference fares even worse than fear, and hatred yet worse again. To be indifferent toward someone is, among other things, to be disinclined to take the fact that they are counting on you to be a reason for acting as counted on. The indifferent simply do not care about responding positively to dependency. Hatred, at least if we accept Hume's analysis according to which it "produces a desire of the misery and an aversion to the happiness of the person hated,"[8] will tend to reverse the valence of the consideration that someone is counting on you, so that instead of taking it to be a reason in favor of acting as counted on, dependency is seen as an opportunity for harm or exploitation. It is partly constitutive of both hatred and indifference that they block the right kind of responsiveness to the fact of another's dependency; hence they are rightly widely taken to be incompatible with trustworthiness.

The simple story of trustworthiness, which focuses on the role played by the consideration "he's counting on me" in the trustworthy agent's deliberation, 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

[8] David Hume, *A Treatise on Human Nature*, ed. L. A. Selby-Bigge, bk. 2, pt. 2, sec. 6 (Oxford: Clarendon, 1978), 367.

have goodwill toward the truster: just in virtue of being positively 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.[9]

Now to the task of adjudicating 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."[10] 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 toward the truster in their deliberation and action. 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 counted on to advance, for trust and trustworthiness are always tacitly limited in domain. The key to maintaining an ongoing relationship is to meet the expectations that the other party has for that relationship. Ignoring interests outside that area is less likely to jeopardize the relationship than is busybody 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."[11] And—drawing out the implicit account of trustworthiness contained in these remarks—my trustworthiness is my capacity to recognize that my interests are dependent on responding to your (success-critical) expectations.[12] In other words, my

[9] I make this mistake in my "Trust as an Affective Attitude." Since accounts of trust are typically taxonomized by the motivational structure that they impute to the one trusted, this is no small mistake.

[10] Hardin, *Trust and Trustworthiness*, and Philip 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 partly constituted by dispositions to take the fact that the other is counting on me as a reason to act as counted on. If we think our amiability is being exploited, it quickly fades. For an 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 (2008): 237–54.

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

[12] See ibid., 28, where Hardin makes it explicit that the key is responding to what "one is trusted to do."

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 that self-interest stands to responsiveness to dependency in roughly the same way that fear does. Unlike indifference and hatred, self-interest is not constitutively incompatible with trustworthiness, but like fear and unlike goodwill or conscientiousness, the responsiveness of the self-interested requires the presence of an additional incentive. Because of this, it is unstable as a motive for responsiveness. We can see this by looking at what happens when self-interest and responsiveness to dependency come apart. Hardin's own example, drawn from *The Brothers 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.[13] When the officer's posting ends, Trifonov refuses to return the borrowed money and disavows the existence of their "arrangement." Without the added incentive of self-interest, responsiveness withers; hence self-interest is fundamentally unstable as a motive for trustworthiness.

III. THREE-PLACE TRUSTWORTHINESS

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. Moreover, the focus has been exclusively on that most controversial part of any account of trustworthiness, the motivational structure of the trustworthy, where I advocated maintaining neutrality in favor of a focus on the distinctive reason to which the trustworthy are responsive. The discussion has thus entirely left out reference to competence. But the conceptual role argument also highlights the importance of being able to harness the power of the variety of human competences to achieve things we cannot achieve alone. It is time to bring competence back into the picture and to add the missing precision.

Let's begin with a canonical statement of what I'm going to call "three-place trustworthiness," so as to bring out its structural parallel with trust, which is universally acknowledged to have tacit three-place structure:

> Three-place trustworthiness: B is trustworthy with respect to A in domain of interaction D, if and only if she is *competent* with respect to

[13] Ibid., 1–3.

> 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 *compelling* reason for acting as counted on.[14]

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 fulfill 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. In assessing whether an agent has taken a consideration to be compelling, background norms for when an excuse counts as good enough are implicitly called on. These norms can be highly local—understandings among a particular group of high school students, for example—or they can be assumed to have broader application and, in some contexts, can even be moral norms. 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 for acting as counted on, 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 behavior 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.[15] 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 we might refer to the fact of the other's dependency. We might express it to ourselves or to others in

[14] I am going to work with the "domain of interaction" formulation, rather than the more popular "with respect to action Z," because I think three-place trustworthiness has a certain "breadth" and thus must extend beyond the performance of a specific action or action type. 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 can be made for it, as I did in my "Trust as an Affective Attitude." However, nothing significant hinges on whether three-place trustworthiness is formulated in domain or action terms, so I do not reengage that debate here. Readers are free to use their preferred formulation.

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

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.[16] 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 with respect to A in domain D 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 with respect to A in D, 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 latter 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

[16] Nevertheless, trustworthiness is not always a matter of trust responsiveness. A discussion of what partner concept of trust that 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 Sec. II that we can count on things as well as people). There is a family of accounts of trust that could fit this account of trustworthiness, including a version of my earlier account, corrected for the mistaken focus on goodwill. Trustworthiness, as requiring responsiveness to the fact of the person’s counting on you, can thus be shown where the other does not (or does not yet) trust. (People can count on you, without counting on you taking the fact they are counting on you into account.) Take the now familiar example of Kant and his neighbors who use his regular habits to tell the time. Kant probably is trustworthy with respect to providing the time to his neighbors 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 neighbors 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.

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 three-place trustworthiness suggests strategies for promoting it. There are two different kinds of strategies that can be pursued. First, we can work on increasing the prevalence of those motivational structures that constitutively enable responsiveness to dependency in a given context. 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 are three-place trustworthy with respect to many domains in their interaction with many 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 three-place trustworthy toward many in many domains, 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 three-place trustworthiness of the almost trustworthy, and it may not take much to tip them over the line into trustworthiness proper. There is a key difference between these institutional and interpersonal strategies and strategies that aim to add extra incentives for outward compliance with others' expectations, such as fear of penalty. Strategies of this latter sort produce behavior that only mimics that of the trustworthy and that is vulnerable to shifts in incentive. Further, adding external incentives can sometimes erode internal motivation to trustworthiness, such as comes from, for example, conscientiousness. In contrast, when we reduce the field of competition from reasons of self-interest, we increase the likelihood of actions that spring from responsiveness to the dependency of others and thereby display genuine trustworthiness.

IV. RICH TRUSTWORTHINESS

The notion of three-place trustworthiness, indexed as it is to specific domains and parties, is something of a term of art—a notion we need to introduce to talk about a property that matches trust once we recognize trust's three-place structure. It might be objected, however, that *any* three-place account of trustworthiness—and not just the one proposed here—is missing something, for we often apply the word "trustworthy" to people without domain qualification, and we would certainly withhold it from someone who responds to the dependency of a select few individuals across a meager range of domains, even though the three-place analysis yields the verdict that they are trustworthy with respect to those people in those domains. Nor would we typically return a verdict of "untrustworthy" for someone who failed to respond to an unreasonable dependency that they did their best to ward off. Yet, it follows from the three-place analysis that they are untrustworthy with respect to that person in the domain in question.

We can get clearer about what is missing from three-place trustworthiness by digging deeper into the conceptual role argument. We not only want that there be people out there who will enable us to extend the efficacy of our agency by doing on our behalf something that we cannot do, or do as easily or as well, for ourselves. As finite agents, we have limited information and limited time to search for more. We want the competent who can be counted on in the ways we need to identify themselves, and we want those who are not up for a particular form of dependency, whether because they lack the competence or the inclination, to identify themselves before we count on them in ways that are apt to be disappointed. We want those we can trust regarding a particular domain to signal their trustworthiness to us, so we can work out where—and where not—to turn.

Those who are willing to signal where they are and are not three-place trustworthy take on some of the burden of helping trusters place their trust wisely. Insofar as they are willing to take on this burden, they show a distinct kind of responsiveness to the inevitable fact of human interdependency. They recognize that we finite beings must sometimes depend on others to help us work out where we may count on responsiveness. They therefore have a further strand to their recognition of the fact of human dependency than do those who are merely three-place trustworthy. For this reason, call them "richly trustworthy."

Rich trustworthiness is two-place in structure: B is richly trustworthy with respect to A just in case (i) B is willing and able reliably to signal to A those domains in which B is competent and will take the fact that A is counting on her, were A to do so, to be a compelling reason for acting as counted on and (ii) there are at least some domains in which B will be responsive to the fact of A's dependency in the manner specified in i.[17] Rich trustworthiness admits of degrees, both in capacity reliably to signal and in the range and importance of domains in which responsiveness to dependency obtains. There is no precise cutoff point below which

[17] Could there be something in between three-place trustworthiness and rich trustworthiness, as I have defined these notions? Could, for example, an agent's willingness to signal be domain limited? I don't see why not. However, two-place rich trustworthiness is the more interesting property, and the one that we have an investment in seeing people instantiate.

we would say that someone is missing three-place trustworthiness in too many domains to count as richly trustworthy whatever their signaling competence. Nor are there rules regarding when someone may signal their way out of responsiveness to a dependency without this counting against their rich trustworthiness. At most, there are locally negotiated understandings.

When we talk about cultivating trustworthiness, we sometimes have three-place trustworthiness in mind, as when, for example, we talk about ways of fostering the trustworthiness of doctors with respect to their patients. However, often, when we talk about cultivating trustworthiness, our target is rich trustworthiness: we want both to increase the range of domains over which people will be competent and responsive to dependency and to improve those capacities required to have a reliable grasp of these zones of competence and to be able to signal them to others.

Someone might have three-place 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: even though they reliably signal who can count of them for what, some potential trusters are 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 three-place trustworthiness, they lack rich trustworthiness. 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?'"[18] We want them to signal their trustworthiness in a domain. One of the best ways to do this is 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. But there are other ways, too, including by showing that they are relevantly similar to others who can be trusted in this domain. We want them reliably to signal in, so that we can identify those whose agency is recruitable to extend the effectiveness of our own. We also want them reliably to signal out if they do not have the competencies on which we might base potential dependencies or if they are not willing to be responsive to those dependencies. Unsignaled or unreliably signaled three-place trustworthiness is no use to us.[19] Rich trustworthiness involves both willingness to signal in and to

[18] Baier, "Trust and Anti-trust," 244.

[19] 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).

signal out, but it is no mere truth in advertising, as if one could count as richly trustworthy by always correctly signaling that one would not be responsive to any dependencies at all.

Rich trustworthiness requires capacities significantly more sophisticated than those required for three-place trustworthiness (which are themselves not trivial). Correctly signaling my trustworthiness (to a person regarding a domain) requires grasping what the other will count as a signal. Signaling 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 signaling what we can be counted on for. Individual competence in signaling requires understanding what is being signaled to whom through these socially mediated "standing channels." Because we live in a world in which how we present ourselves and who we are taken to be carries with it social meanings, we are inevitably signaling, rightly or wrongly, who can count on us for what. In order to signal, we need do nothing at all. Sometimes signaling correctly requires merely that we do not disrupt these standing social signals. At other times, signaling correctly requires active moves to disrupt them whether by signaling that we can be counted on for less than might have been reasonably supposed, given who we are taken to be, or by signaling that we are reliable in ways that might not be expected of us. Standing social signals can be exploited by untrustworthy agents, whether actively by signaling they have the properties characteristic of agents who tend to be trustworthy in a given domain when they do not, or passively by allowing signals that they know to be false to stand uncorrected.

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 ongoing reflective self-monitoring of my own competences so that I know them and their limits.

Though rich trustworthiness is harder to cultivate than three-place trustworthiness, there is much that we can do to scaffold it in ourselves and in others. 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 signaling, 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 richly trustworthy and through our ongoing interaction with them that we are sustained in our trustworthiness as our trustworthiness receives uptake in trust. It is in part because we have the concept trustworthy that we become trustworthy.

I have defended an account of three-place and rich trustworthiness that is open with respect to the motivational structure of the trustworthy. This motivational openness 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 the availability of three-place trustworthiness and our ability to identify those who have it. Better, goes a common view, to economize on trust, since trustworthiness can be predicted to be in short and hard to identify 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. 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 economize on trust before our cash reserves of trustworthiness run out.

The accounts of both three-place and rich trustworthiness are more optimistic. Though there is no denying that face-to-face societies will make motives that support three-place trustworthiness more prevalent, they are not necessary conditions for it. One needs neither goodwill (except in the minimal sense associated with responsiveness itself), nor ongoing relationships, nor even shared values to be trustworthy. 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.

V. 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. In this section, I argue that the classical lists are correct in excluding trustworthiness, but not quite for the reasons contemporary theorists who are skeptical of its being a virtue have so far identified. Some contemporary arguments against trustworthiness as a virtue confuse three-place trustworthiness with rich trustworthiness; others assume too simple a model of the virtues and, if sound, would count equally against uncontroversial virtues such as honesty. The case against trustworthiness as a virtue needs to be reprosecuted, which is my goal in this final section. Looking at the question of whether trustworthiness is a virtue enriches our understanding of the interest that we, as finite, reflective, social beings have in a nonmoralized conception of trustworthiness such as the one I have defended here. On the account defended here, trustworthiness is decoupled from the normative worth of the expectations that it meets.

When asking whether trustworthiness is a virtue, we first have to clarify what we mean—are we asking whether three-place trustworthiness or rich trustworthiness is a virtue? The only coherent question here is whether rich trustworthiness is a virtue. Recall that a virtue is a stable state of character that is an excellence. Three-place trustworthiness is not a stable state of character, nor is it an excellence or its lack a deficiency; hence it is not even a candidate for being a virtue. It is not well formed to say that someone has or lacks three-place trustworthiness, since it must always be ascribed to someone with respect to a person (or group of persons) and regarding a domain of interaction. A can have three-place trustworthiness with respect to B in a domain of interaction D and yet this trustworthiness not extend to others in like domains of interaction or to B in other domains. Further, the three-place trustworthiness that, say, a hit man shows to those who employ him is no excellence. Finally, failures of three-place trustworthiness—even failures in the service of good ends—do not necessarily involve falling short of excellence. Finite human agents have patchworks of competence and so our three-place trustworthiness is always limited. Even when we are generous in our responsiveness to the dependency of others, the limits of our competence mark the limits of our three-place trustworthiness. One is not falling short of moral excellence if one lacks the competence to be trustworthy with respect to many domains, from plumbing, to medicine, to finance.

Rich trustworthiness, involving as it does the ability to monitor and signal one's competences, is not similarly constrained by the limits of competence. I can be richly trustworthy despite large gaps in my competences, so long as I know the location of those gaps and take measures to head off dependencies in areas where I am not able to follow through.[20] Whereas basic trustworthiness clearly fails to meet the requirement that a virtue be a stable state of character that is an excellence, rich trustworthiness—when held in exemplary measure—just might meet it. Rich trustworthiness is the correct target of any investigation into whether trustworthiness is a virtue.

A. *The Standard Case against Trustworthiness as a Virtue*

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

1. There need be no fault in refusing to respond to unsolicited trust with trustworthiness, for sometimes trust can itself be an imposition.[21]
2. Trustworthiness can be in the service of bad ends as well as good ones. Evil thrives when evildoers work together.[22]
3. One can be required to respond to trust, extended in service of evil ends, with "trust busting."[23]
4. It is not always wrong to actively elicit trust and then "bust" it with treachery.

I begin with the first charge: prima facie, there is a difference between trustworthiness and other recognized virtues such as benevolence in that, within the limits of justice and capacity, we are required to respond to need with benevolence. We do not get let off the hook by saying we would rather not. In seeming contrast, one need not respond to trust with trustworthiness. Sometimes, by their trust, others can attempt to manipulate you into responding to their dependency, and it need be no fault on your part if you refuse to succumb to their pressure.

[20] Ibid., chap. 1, 1–34. 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 to those who shared similar values.

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

[22] Baier, "Trust and Anti-trust"; Amy Mullins, "Trust, Social Norms, and Motherhood," *Journal of Social Philosophy* 36 (2005): 316–30.

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

Having distinguished rich trustworthiness from three-place trustworthiness, the rejoinder to this objection to trustworthiness as a virtue is obvious: the person who refuses to respond to the fact that someone is counting on them in a domain fails to display three-place trustworthiness toward them in this matter. But it does not follow that they must fail to display *rich* trustworthiness—the richly trustworthy signal who can count on them for what and so they do not merely turn their backs on poorly placed or presumptuous trust. They will indicate that it is misplaced and invite it to be withdrawn.[24] No failure, then, of rich trustworthiness, provided the limits of the dependencies one will be responsive to have been appropriately signaled. The first consideration against trustworthiness as a possible virtue misfires, as it targets only three-place trustworthiness, which never was a serious candidate for being a virtue.

The second charge against trustworthiness as a virtue is that it can further bad ends as well as good ones. This time the objection is on target—rich trustworthiness can indeed further bad ends—but it is weak. There are clearly recognized virtues about which a similar complaint can be made, and recognized ways of responding to such complaints. For example, 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 (rich) trustworthiness must, in fairness, be allowed the same defense.

The doctrine of the unity (or necessary compresence) of the virtues grounds this rejoinder. According to the doctrine of the unity of the virtues, virtues must come in packages. The doctrine gains support from the assumption that virtues are states of character that issue in only right action together with the observation that there are situations in which different virtues appear to pull the agent in different directions. Justice and kindness can appear to conflict: if one's kindness is not to result in wrong action, one must also possess the virtue of justice, since kindness that leads to injustice is no real kindness. A virtue is, among other things, a sensitivity to reasons of a certain kind, but given that there are no limits on the combinations of virtue-relevant reasons that a situation can present, such apparent conflict between the virtues is always a possibility. Thus, if the virtues are to result in only right actions, one cannot have one without having them all. The virtuous must have the practical wisdom to discern which consideration is morally salient in which circumstances.[25]

The doctrine of the unity of the virtues also provides a way of responding to the third and fourth objections, which are related. According to the third objection, one can be required to "bust" trust extended

[24] Potter, *How Can I Be Trusted?*, 26–27.

[25] For this formulation of the Socratic argument, see John McDowell, "Virtue and Reason," *Monist* 62 (1979): 331–50, 331–33.

in the service of bad ends. The objection needs unpacking: by trust busting must be meant committing acts of treachery that blow apart corrupt trust relationships. Thus, objection 3 claims one can be required to be untrustworthy, though it remains unclear whether it is rich or three-place trustworthiness that is at issue here. If trustworthiness (of the kind in question) is a virtue and hence untrustworthiness a vice, this claim has a decidedly odd ring, for surely one cannot be required to exhibit a vice. Nothing can require one to be cowardly, dishonest, or cruel. The fourth objection ratchets the third up a notch by clarifying the ambiguity in the framing of the third objection. Perhaps the agent confronting an apparent requirement to be untrustworthy has gotten herself into this lamentable situation through a failure of rich trustworthiness—perhaps she failed to signal that she was not the kind of person to become involved with a project such as this. The fourth objection points out that one can be required to wrongly signal what one can be counted on for and then deliberately fail to carry through: rich trustworthiness is unquestionably the target of the fourth objection.[26]

Once again, the doctrine of the unity of the virtues comes to the rescue. To see how it grounds a rejoinder, consider, as a parallel, a recognized virtue such as honesty. The honest can sometimes be required to lie, but this is not the same as being required to exhibit the vice of dishonesty. For example, 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 even actively seek a reputation for honesty among the enemy in order that their lies might be more readily believed. Thus, the honest might be called on to behave in ways the third and fourth objection suppose undercut the claim that trustworthiness is a virtue, yet no one doubts that honesty is a virtue. The virtue of honesty consists not merely in a disposition to tell the truth, but rather, among other things, in a disposition to recognize the importance of taking "it is the truth" as a reason in practical deliberation. Understanding its importance includes having the practical wisdom to discern when that consideration is less important than the fact that justice requires intervening to stop wrongs. A lie correctly told in the name of justice does not count against the teller's honesty. Likewise, the defender of trustworthiness's claim to be a virtue will say that letting down someone who is counting on you in the name of justice no more counts against your possessing the virtue of trustworthiness than telling a lie in the name of justice makes you dishonest. It is one thing to be required to lie, or to let people down; it is another thing entirely to be required to be dishonest or to be untrustworthy.

It looks like the case against trustworthiness as a virtue has failed: either it targets three-place trustworthiness, which never was a real candi-

[26] Hereafter, I will use "trustworthiness" to mean rich trustworthiness, since it has to be the target for those arguing both for and against trustworthiness as a virtue.

date for being a virtue, or it ignores the need for the virtues to come as a package. Given that—at least for all that has been so far—trustworthiness behaves in the same way as other recognized virtues such as courage and honesty, it seems it can keep the title of virtue, if they can. The case against trustworthiness as a virtue must be either abandoned or reprosecuted.

B. Reprosecuting the Case against Trustworthiness

Current arguments against trustworthiness as a virtue do not go deep enough into the nature of trustworthiness itself to be able convincingly to separate it from other character traits that are recognized as virtues. We can get an inkling of difference between trustworthiness and recognized virtues such as honesty by considering the following four 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.
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.

Both 1 and 2 make sense and show honesty and trustworthiness appearing to behave the same. They are instances of the insight contained in the doctrine of the unity of the virtues: a lie correctly told in the name of justice does not impugn the teller's honesty; letting down someone who is counting on you because practical wisdom reveals that, in the circumstances, considerations of justice are stronger does not count against your trustworthiness. The difference between honesty and trustworthiness comes into view when we consider propositions 3 and 4: 4 makes sense in a way that 3 does not. Though other virtues, such as justice or compassion, might sometimes call for us deliberately to tell a falsehood (as in the case of the spy infiltrating the genocidal army), honesty takes as its signature reason "it's the truth," and honesty alone speaks for a solution that respects the truth. Other recognized virtues behave as honesty: we can be required by justice to deal harshly with a bully, but we cannot be required to cause distress to the bully out of compassion to her victim. Compassion alone speaks for a solution that will relieve the distress of the victim without causing further distress. If there is no such solution, then justice invites us to think of the bully's distress as less important than the distress they are unjustly inflicting on their victim. None of this is to say that there cannot be real moral dilemmas, or that it can sometimes be difficult to determine (or even indeterminate) what a virtue requires of us, let alone what is overall virtuous.[27]

[27] For the notion of "overall virtuous," see Christine Swanton, "A Virtue Ethical Account of Right Action," *Ethics* 112 (2001): 32–52, 45–48. My remarks here are not meant as a definitive account of the signature reasons of the virtues of honesty and compassion, only as plausible approximations. Those who prefer alternative accounts are

Compassion, honesty, and the rest are not hostage to the expectations of others. Trustworthiness, on the account of it that I have defended here, is. Trustworthiness's signature reason is "he's counting on me," but all sorts of people can count on you for all sorts of incompatible things, things which cannot, even in principle, be mutually realized. Trustworthiness as, inter alia, a disposition to be deliberatively responsive to "he's counting on me" has the potential to be inwardly riven in the way that honesty, kindness, and justice do not. As a result, trustworthiness is capable of conflicting with itself so that trustworthiness can require one to be untrustworthy, or so I will argue.[28]

Let's look a real-life example. Suppose that I have become deeply involved in two community groups whose values I hold important. One group works on maintaining local historical heritage, including the flavor of the Victorian streetscapes that characterize the neighborhood. The other group promotes and celebrates social inclusion and neighborhood diversity. Suppose I have indicated a willingness to produce materials in support of each group's campaigns. Both groups have come to count on me to do this, and neither group has an alternative way of producing these materials, because, given my past reliability, neither group has had reason to seek a backstop. Now suppose the two groups find themselves on opposing sides in a planning tussle over a proposal for a high-rise development, out of scale and tone with the current streetscape, that will nevertheless do something to make housing in the neighborhood more affordable and thus open to a more diverse group of people. I can find myself being counted on, by two different sets of people, to do two noncontingently incompatible things: to help get the proposal through the council and to help kill it dead. Two different groups are counting on me to further, and to block, the self-same goal. Being trustworthy to one group requires me to let down the expectations of the other, expectations which I have actively created and actively indicated would be met. Trustworthiness with respect to one group requires untrustworthiness of me with respect to the other. Whatever I do, I must let down one group; whatever I do, one group will rightly feel betrayed.[29]

welcome to substitute their own. My only stake is in an account of the signature reason of trustworthiness.

[28] Can't, say, compassion conflict with itself, so that we can be required by compassion to be cruel, as when, say, we cause distress to avert a greater distress? Only if causing distress is incompatible with compassion, which is implausible. The example that follows aims to show it is plausible that trustworthiness (to some) can require untrustworthiness (to others).

[29] It might be objected that the group that loses out shouldn't feel betrayed. Perhaps there has been no breach of trustworthiness because the fact of their dependency is still being taken to be a compelling reason, but it is overridden by the dependency of the other group, as the account of three-place trustworthiness allows. However, from the perspective of either group, given what they are counting on the trustee to do, allowing their dependency to be overridden is itself a betrayal. Thus, the case contrasts with one where the trustee is unable to fulfill expectations because, say, they have to attend to a sick parent.

In its capacity to be riven in this way, trustworthiness differs from all other recognized virtues, with the possible exception of loyalty, about which I'll say more in a moment. Why, however, should we suppose that the potential for conflict within trustworthiness makes it unfit to be a virtue? Start from the plausible intuition that one cannot be required to exhibit a vice. Suppose that trustworthiness is a virtue, then its opposite, untrustworthiness, must be a vice. Hence you cannot be required to be untrustworthy. But you can be required to be untrustworthy as the above example shows; therefore, trustworthiness cannot be a virtue. It is not open to a defender of trustworthiness's status as a virtue to appeal to the doctrine of the unity of the virtues, as they did in reply to the simpler though somewhat related objection (objection 4) considered in Section V.*A*: the conflict isn't between trustworthiness and another virtue, but within trustworthiness itself. Nor can the defender of trustworthiness's status as a virtue say that the argument confuses basic trustworthiness with rich trustworthiness and so misses its target. So long as the expectations are legitimate and not the result of a failure to signal the kinds of things one can be counted on for, then even someone who is richly trustworthy to an exemplary measure can yet find herself in this kind of conflict.[30]

At this point, a defender of trustworthiness as a virtue might appeal to loyalty, for only loyalty appears to behave in the same way as trustworthiness, with the apparently paradoxical consequence that, at least when loyalties conflict, one can be required to be disloyal. If trustworthiness is like loyalty and unlike, for example, honesty, courage, and justice, then that is bad news for the claim that trustworthiness is a virtue, given that loyalty's hold on current lists of the virtues is contested.[31] But things are even worse. If loyalty is to keep the title "virtue" in 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.[32] Our commitments can conflict contingently, and we must

[30] None of this is to say that there aren't better and worse—or even more or less trustworthy—ways of going forward. It will be better for me to clearly explain to both groups the reasons for my decision about which to support on this occasion. It will also be important to signal to both groups my limits and what they may and may not count on me to do in the future. These are both important ways of repairing the breach in trust that my actions caused. The point here is only that I have, indeed, let down those who, given my past signaling, reasonably counted on me, and to that extent have failed to be richly trustworthy to one group in order that I might be richly trustworthy to the other.

[31] For an argument against loyalty as virtue, see Simon Keller, *The Limits of Loyalty* (Cambridge: Cambridge University Press, 2007), 144–62.

[32] 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.

choose 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 rather than in our own commitments. The ground of loyalty lies in our commitments and the demands they impose, leaving us the option of altering the demands by altering the commitments. The ground of trustworthiness, on my account, lies in the expectations of others and the demands they impose, release from which is not up to us. On my nonmoralized account, these expectations can be for help furthering ends that are good, bad, or indifferent.

None of this is to deny that we could, if we wanted to, define up a fully moralized conception of trustworthiness, such that there can be, by definition, no brotherhood of thieves knit together by in-group trustworthiness. However, the conceptual role argument suggests that there is good reason to resist this moralizing move. We are finite dependent social beings, and we are less, often much less, than fully virtuous. We want there to be others who will be responsive to our counting on them so that we can extend the efficacy of our agency through their help. We want more than this—we want these people to help identify themselves to us, so that we will know with whom there is the potential to get into cahoots, for good, for bad, or for indifferent projects. Rich trustworthiness thus identifies a quality that we have reason to care about, but part of the reason we care about it is because it resists moralization.