Being Pragmatic about Trust

Philip J. Nickel

Eindhoven University of Technology, p.j.nickel@tue.nl

Author ms. Final version published in P. Faulkner and T. Simpson, eds., *The Philosophy of Trust* (Oxford University Press, 2017): 195–213.

1. Introduction

Despite substantial attention to conceptual questions about trust from philosophers and non-philosophers alike, trust remains an ambiguous and contested concept. Scholars wrestle with two problems: the familiar philosophical problem that we cannot find a suitable analysis of trust on which everybody agrees in their intuitions (by analogy to the problem of the analysis of knowledge, say),[1] and the broader problem that the concept of "genuine trust" investigated by epistemologists and moral philosophers is not the same one used in many explanations of social phenomena on the basis of individual behaviour. In this paper I will argue that the second, broader disagreement has implications for the disagreement among philosophical accounts, favouring accounts that are less restrictive and demanding, and that mark genuine explanatory categories.

A so-called Trust Game (early versions in Camerer and Weigelt 1988; Kreps 1990: p. 100; Berg, Dickhaut and McCabe 1995) brings out the conceptual

[1] See Simpson (2012) for a description and diagnosis of the first problem. My approach differs from Simpson's in emphasizing the possibility of pragmatic arguments that may help to resolve these two problems.

discrepancy I have in mind. Here is one version of the game: suppose if Francesca gives George $20, George will use it to earn $80, an opportunity not available to Francesca by herself. However, it is completely up to George whether he returns any part of the money to Francesca. Suppose Francesca is disposed to give George $20 in this situation. As the name of the game suggests, explanations of cooperation (or non-cooperation) in terms of individuals' needs and interests treat Francesca's disposition as trusting without regard to the particular reasons or motivations she might have. Suppose that on this particular occasion of the Trust Game Francesca guesses that George will give her back a portion of the money because she thinks *he believes she is his enemy, and wants to keep his enemies close*. She attributes a cunning motive to him, and for present purposes she is willing to use that to further her own ends. From the point of view of explaining cooperation, there is no reason not to regard this as a potential motive for trust. Strategic, calculative reasons are an important class of reasons for cooperation and can take many forms. They can consist of a general expectation that the two interacting parties will encounter each other many times in the future (so that it makes sense for them to cooperate now); a sheer lack of better options; a desire to protect one's reputation; or they can consist of particular beliefs such as Francesca's.

For most philosophers, such reasons are unsatisfactory as reasons for trust. Intuitively, not just any reason or motivation disposing Francesca to "invest" in George in such a situation is compatible with genuine trust. Intuitively, Francesca's disposition is not trusting even though she has sufficient confidence that George will return some of the money that she is willing to give him. More generally, dispositions toward reliance based on this sort of *strategic expectation* must be different from

dispositions based on trust. As a result, the concept as used in explanations of social phenomena and the philosophical concept clash in what they treat as trust.

To spell out the narrower, intuitive concept, philosophers often introduce conceptual restrictions on the allowable motivations and reasons embodied in a person's disposition to rely on another person, if it is to count as trust. On various accounts, a trusting person must have:

- Optimism that the person relied upon will act with competence and goodwill (Baier 1994: p. 98; Jones 1996) or moral integrity (McLeod 2002);
- An affective expectation that the person will be responsive to the fact that one is depending on her (Faulkner 2007);
- A normative or moral expectation that the person will perform a certain way (Nickel 2009), for which we hold them responsible (Walker 2006: p. 78);
- A belief that the person relied upon is trustworthy (Hieronymi 2008).[2]

These restrictions imply, more or less, that Francesca's strategic expectation of reliability does not count as genuine trust. Her disposition to invest money in George is not based on optimism about George's goodwill or integrity, nor on an affective expectation that George will be responsive to her reliance, nor on a moral expectation, nor does it involve a belief that he is trustworthy (understood the way these notions are intended). I will come back to some of the details later, but what I want to emphasize at first is what these views *share*—their focus on conceptually restricting the motives and reasons of trust. I will label them *restrictive* views of trust,

[2] "One person trusts another to do something only to the extent that the one trustingly believes that the other will do that thing" (Hieronymi 2008: p. 214).

and the alternative the *unrestrictive* view. I will also speak of less restrictive and more restrictive views.

Considerable philosophical motivation of various restrictive views has been provided in some of the articles cited above. In this chapter I explore what might be said on behalf of *un*restrictive and less restrictive views on which the concept of trust is left open, potentially covering all manner of dispositions to rely on another person. Russell Hardin (2006) formulates the point I want to develop, suggesting an *argument from explanatory potential* according to which less restrictive views of trust are more useful because they can provide a more adequate explanation of the origin of cooperative behaviours and social institutions than more restrictive views.

I begin by presenting the argument from explanatory potential. I contend that the degree to which restrictive views are undercut by this argument depends on exactly what they require of the trustor. I consider in somewhat more detail how two of the more promising restrictive accounts fare. At the end of the paper, I briefly consider the reasons for thinking that such an argument actually might partly determine our concept of trust.

2. Trust in an Explanatory Role

In this section I develop the argument from explanatory potential and apply it to restrictive and unrestrictive theories of trust. The basis of the argument is a methodological constraint on theorizing about trust: trust should (a) be explained as the outcome of central concerns or interests of the relevant actors, and (b) explain the emergence and sustenance of cooperative practices and social institutions

(Hardin 2006: p. 16).[3] In what follows I will call this the Explanatory Constraint, and I will call (a) the Input Condition and (b) the Output Condition. It is explicitly a constraint about *social explanation*, and is meant to allow for empirical research on trust that explores the particular mechanisms and types of causation involved. It is neutral about other possible explanatory aims (e.g. investigating the developmental significance of a trusting orientation for the individual, or philosophically motivating the linkage between trust and certain moral attitudes).

[3] Hardin's formulation is that it should "be explained . . . as the outcome of behaviors guided by some central concern or motivation of the relevant actors" (Input Condition) and should yield "explanations of behavior and social institutions" (Hardin 2006: p. 16) (Output Condition). In my formulation I have dropped the idea that trust should be the outcome of behaviours, since it seems that it might also be the simple result of the concerns and interests themselves. Hardin does not develop the argument in any detail.

The claim that trust explains the emergence and sustenance of cooperative practices and social institutions is not meant to imply that it is the only factor that explains the development of cooperative practices, nor that cooperative practices conceptually require trust. In fact, Hardin and his colleagues explicitly restrict the concept of trust to dyadic relationships in which one has specific beliefs about the other's motive, and at the same time try to "make sense of a wide array of devices for organizing cooperative behavior in the absence of trust or in the presence of very weak trust" (Cook, Hardin, and Levi 2005: p. 7). Many other factors also play a role on their account. But if the Explanatory Constraint is true, trust has an important role, not only when it is "successful" in the sense that the person one trusts actually performs, leading to a pattern of cooperation or a relationship of reciprocity, but also when it fails and leads to other strategies for securing cooperation, as it sometimes appears to.

In order to see why the Explanatory Constraint is attractive, it is worth considering the situation that would obtain if it were not satisfied. Trust is widely regarded as an important concept for understanding social phenomena. But if trust is not used to describe how central concerns or interests of interacting parties lead to social behaviours, practices, and institutions, then it cannot be particularly important for understanding social phenomena. For example, if a person's needs for food and shelter, and the social opportunities for the realization of these needs, do not normally lead to a further state described in terms of trust—whether fulfilled by the performance of the other, or not—that mediates the formation of cooperative practices and social institutions, then trust cannot play a significant role in explaining how social phenomena emerge from basic human needs. In that case, trust will not be especially important in understanding social phenomena. This argument shifts the burden to those who do not evidently satisfy the Explanatory Constraint, to show why trust nonetheless deserves a place in social philosophy, or to explain why its apparent importance is illusory.

One of the foremost advocates of the unrestrictive view of trust is the sociologist James Coleman. On Coleman's version of the view, trust is simply a disposition to rely on another person in order to satisfy substantial needs or interests, so that the expected gains justify the losses (1990: p. 99). Trust is explained in terms of the needs and interests of individuals (or sub-social groups), together with the opportunities that reliance on others can provide toward fulfilling those needs and interests. The trusting disposition consists of a posited mental state that takes these factors into account and resolves action in light of them. In turn, the successes and failures of actions taken in accordance with such dispositions explain (i) the emergence of stable patterns or practices of cooperation; and (ii) the evolution and

design of norms and institutions that help change the balance of incentives so that failures of trust can be avoided in the future to a greater degree. (In what follows I will sometimes refer to (i) and (ii) together as "practices and institutions.")

Coleman's theory emphasizes the strategic rationality of individuals when it comes to explaining why practices and institutions emerge (see Coleman 1990: pp. 13–19). His attribution of self-interested, strategic rationality to individuals or sub-social agents is an interpretive posit that makes general sense of human behaviour for explanatory purposes.[4] It is elastic in what it treats as rational, allowing rational agents to take, for example, internalized social norms as reasons (Coleman 1990: pp. 292–3). The positing of rational choice is not meant as a conceptual requirement that distinguishes between trusting reliance and non-trusting reliance: the notion of rational choice does not restrict what counts as falling under the concept of trust in the same way that the criteria from restrictive views are meant to. In particular, the unrestrictive view of trust is *not* to be thought of as a conceptual view on which a trusting person always tries to satisfy her individual, self-interested preferences. Such a view would encounter the empirical problem that in one-off variants of the Trust Game, experimental participants often exhibit a disposition to cooperate (Johnson and Mislin 2011). This cannot easily be explained in terms of individual, self-interested preferences.

A possible worry one might have about such a view is that it is *too* broad to allow for empirical explanation. It is not empirically explanatory to say that trust is simply whatever disposition leads one to rely on others: this is like saying that desire is simply whatever it is that leads one to intentional action, or that food is simply

[4] For more on this idea, see Buchak (2016) for her notion of "interpretive decision theory".

whatever nourishes us (in Aristotelian fashion). To say this is merely to label a phenomenon, rather than to explain it. However, the unrestrictive view does not aim to make a *definitional* statement that labels all possible grounds of attempts at reliance on others as instances of trust. Rather, the point is to take a pragmatic view, leaving open what might count as instances of trust so that a range of possible motivations can potentially fit, instead of defining them away analytically a priori.[5]

Under a single concept, so to speak, the unrestrictive view allows for the emergence of stable patterns of cooperation from both strategic and non-strategic dispositions toward reliance. For example, Francesca's strategic reliance, if successful, could lead toward a stable cooperative relationship between George and her; or if it fails, it could lead Francesca to seek to adopt norms and/or institutional controls that would give George (or others like him) more reason to perform (see Figure 12.1). If one of the interacting parties happens to have non-strategic reasons for reliance on the other (e.g. moral reasons), this also counts as trust.

As this example illustrates, the Explanatory Constraint seems to favour an unrestrictive view of trust. If a trusting disposition toward reliance on others can be held prior to cooperative practices and institutions, and is compatible with a broad range of cooperation that takes place within those practices and institutions, then it can explain the emergence and sustenance of those practices and institutions in terms of trust, satisfying the Output Condition. An unrestrictive account has an easier

[5] Thanks to Tom Simpson for prompting me to clarify this. On Simpson's own view, *Ur-trust* is a bit like this (2012). Simpson emphasizes the human need for some such notion in order to ground cooperation and shared activity, and also doubts the prospects of a sharper philosophical analysis.

time doing this than a restrictive view. Because of higher uncertainty about the parties on whom one could rely, and the absence of practices and institutions for guidance in decisions about reliance, the dispositions toward reliance that one finds in situations prior to practices and institutions often include purely strategic expectations of reliability. In addition, it is hard to see how optimism about the other's goodwill, judgements about shared values, and so on—conditions that are constitutive of trust on the restrictive view—could be justified in such conditions.

A different kind of example will help develop the argument further. Consider a situation that approaches a "state of nature": one of the first encounters of European explorers of North America with Native American tribes. The French explorer La Salle and his party approached an Illinois settlement by canoe:[6]

> At nine o' clock, doubling a point, [La Salle] saw about eighty Illinois wigwams, on both sides of the river. He instantly ordered the eight canoes to be ranged in line, abreast, across the stream . . . The men laid down their paddles and seized their weapons; while, in this warlike guise, the current bore them swiftly into the midst of the surprised and astounded [Illinois people]. The camps were in a panic. Warriors whooped and howled; squaws and children screeched in chorus; some ran in terror, and, in the midst of the hubbub, La Salle leaped ashore, followed by his men. None knew better how to deal with Indians; and he

[6] Hobbes claims that Native Americans in his time lived in a state of nature (*Leviathan* I, 13: Hobbes 1968 [1651]: p. 187). My focus is instead on a context of interaction that lies outside shared practices, norms, and institutions. I assume that both parties are internally governed by such practices, norms, and institutions.

> made no sign of friendship, knowing that it might be construed as a token of fear. His little knot of Frenchmen stood, gun in hand, passive, yet prepared for battle. The Indians, on their part, rallying a little from their fright, made all haste to proffer peace. Two of their chiefs came forward, holding out the calumet, while another began a loud harangue, to check the young warriors who were aiming their arrows from the farther bank. La Salle, responding to these friendly overtures, displayed another calumet; while Hennepin caught several scared children and soothed them with winning blandishments. The uproar was quelled, and the strangers were presently seated in the midst of the camp . . . Food was placed before them; and, as the Illinois code of courtesy enjoined, their entertainers conveyed the morsels with their own hands to the lips of [La Salle's party]. (Parkman 1983 [1878]: pp. 836–7)

The unrestrictive view clearly counts the Illinois' willingness to allow La Salle into their camp, and La Salle's acceptance of food from the Illinois, as instances of trust, whereas the restrictive view has a harder time doing so. Suppose that as a result of this initial contact, a cooperative practice of repeated interaction results, such as a continued pattern of interaction in which each party performs in a way that is useful to the other. Or counterfactually, suppose that after the failure of the first interaction an external condition is imposed in order to change the terms of future interactions, such as the "friendly" taking of hostages by both sides (a common practice at the time), or a peace treaty between the French and the Illinois carrying clear sanctions for violations of its terms. The unrestrictive view explains these outcomes in terms of trust, and the restrictive view cannot do so. On restrictive views, both the initial interaction and the patterns of cooperation that result are in all likelihood too

strategic to count as instances of trust. Hence trust does not play a role in explaining whatever cooperative or non-cooperative outcomes result.

We can pose this as a dilemma for restrictive theories of trust. The first horn is to hold that these reliant dispositions prior to applicable practices and institutions do not count as instances of trust. If this horn is taken, then the restrictive notion of trust fails the Explanatory Constraint because trust does not play a role in explaining the emergence of cooperative practices and institutions in such situations. There will be other contexts where it does play such a role, but these will be limited to situations in which there is already a background of shared values or a history of interaction that can provide additional reasons for trust. To accept this horn does not, of course, imply that there is no explanation of the emergence of cooperative practices and institutions in terms of dispositions to rely. It just means that trust does not figure in the explanation. The reason this is a problem is that the underlying process by which cooperative practices and institutions form seems indifferent to whether the conceptual restrictions are fulfilled or not. Whether a person with a disposition to rely on another person is optimistic about their competence and goodwill, for example, seems to make no difference to whether their reliance leads to stable practices of cooperation or to the evolution and/or adoption of norms and institutions that bolster cooperation. It makes no practical difference, in the sense of drawing an important distinction between two different explanatory situations. Hence these restrictions are unmotivated from an explanatory perspective. There is just one phenomenon here, and it appears to be the broader one referred to by the unrestrictive view. (We will have reason to revisit this below when we consider some specific restrictive accounts.)

The second horn is to hold that despite all appearances, people such as La Salle and the Illinois tribespeople often do trust (in the restrictive sense) in such stark situations, prior to practices and institutions. Here we can still explain the emergence of practices and institutions in terms of trust, but only at the cost of holding the implausible view that people in such unpredictable situations have the special motives that restrictive views require for trust. The second horn creates two serious problems for restrictive views. The first is that it is not plausible on the face of it that the Illinois tribespeople, La Salle and his party, or many others in comparable situations, satisfy the restrictive conditions on trust (by having optimism about the goodwill of the other, belief in their trustworthiness, affective expectations of their responsiveness, or moral expectations of them). This is not how their attitudes are described in the account. Hence the explanation involves an implausible attitude ascription.

The second problem with the second horn is that even if this attitude ascription were correct it would require a basic sort of irrationality at the heart of a trust-based explanation of practices and institutions, because there really is not sufficient *reason* to justify such attitudes. Such situations offer only a little by way of moral assurances or common norms that would ground a strong judgement that the other will be trustworthy or take the interests of the other party as intrinsically important. In the Illinois–La Salle meeting, although the two parties have never encountered one another before and share almost no common practices and institutions, a means is found of communicating a sign of reliability: the calumet or “peace pipe,” a ceremonial tobacco pipe the display of which is used as a symbol of non-hostility. Each party possesses such an item, and according to Parkman’s description it is instrumental in establishing initial cooperation. This could perhaps be seen as a

fragile application of a social practice to facilitate cooperation. However, although such a symbol may provide some small evidence of goodwill or trustworthiness or shared values, it does not constitute an adequate (still less conclusive) epistemic reason for any of these things, and its use on one very new occasion does not yet constitute a practice here. Admittedly, each party recognizes that the other is aligned with a larger group of people (the French settlers, and the various subgroups of the Illinois) who might eventually react or retaliate based on how the interaction proceeds. Each party also realizes that it may have many occasions in the future to rely on the other in which it could be useful to cooperate (a factor emphasized by Hardin (2006)). These facts certainly give some reason to rely on the other party, but not of the kind that seems to be assumed by the restrictive views of trust.

This argument does not apply equally to all restrictive views. Different restrictive views are more or less compatible with the Explanatory Constraint, depending on the extent to which they require that a trusting person has very particular emotional or cognitive states. A view like Baier's, on which the trusting person must be optimistic about the competence and goodwill of the other, is less restrictive than a view like McLeod's (2002) on which the trusting person must be optimistic that the other has moral integrity.[7] In order to consider this matter in more detail, in the next two sections I consider two restrictive views: Baier's and Faulkner's. I will not consider views, such as Hieronymi's (2008), that link trust with full belief in the trustworthiness

[7] McLeod situates this claim within a prototype theory of trust on which "our concepts are more malleable than traditional analytic philosophy makes them out to be" (2002: p. 14) but where a certain number of prototypical features must be present in order for something to count as trust to a particular degree. In such a view there may be room for trust in a wide sense to be socially explanatory.

or reliability of the other, nor will I consider McLeod's (2002) view. In my view their requirements are simply too stringent on any reading to satisfy the Explanatory Constraint for the types of cases we are interested in.

3. Baier's View and the Explanatory Rationale for Restrictions

In defence of her own restrictive view of trust, Baier has argued that the Hobbesian conception of trust, taking interactions between strangers as paradigmatic, ignores the experiences of trust within family and intimate relationships. She takes the side of Hume against Hobbes, echoing Hume's famous point against social contract theory that there is no such thing as interaction not yet conditioned by the experience of being raised in a "family-society" (Hume 1998 [1751]: p. 88).

There are two claims here. The first is that interactions between strangers, like the La Salle–Illinois encounter, are not paradigmatic for trust. Apart from the fact that many *have* found these sorts of encounters to be paradigmatic of trust, this first point does not threaten the idea that there are contexts of interaction where the emergence of practices and institutions occurs for the first time, not merely as an extension of existing practices and institutions. We are often interested in explaining the emergence of specific cooperative behaviours and institutions in concrete contexts.[8] We can easily find situations that exemplify this *within* a family society, for

[8] Empirically studied examples include the formation of institutions for administering American college entrance examinations in terms of the needs of elite colleges and universities that began to search for students from a much broader, national pool (Coleman 1990: pp. 647–8), and the design and implementation of reputational feedback mechanisms on eBay (Utz, Matzat, and Snijders 2009).

example in questions over how property of a deceased person should be divided among family members where custom and the law are unclear or nonexistent.

The second point, more important for us, is that the motives of trust are conditioned by our genetic and developmental inheritance, including the experiences of trust and trustworthiness of early childhood. In the context of child development, optimism about the goodwill and competence of another (often the parent) is normal. What Baier calls “infant trust,” the instinctual reliance of young children on their parents to care for them, has a kind of *psychological* priority whose influence endures until adulthood and can be expected to influence people’s behaviour in situations such as the Trust Game.

> That infant trust normally does not need to be won but is there unless and until it is destroyed is important for an understanding of the possibility of trust. Trust is much easier to maintain than it is to get started and is never hard to destroy. Unless some form of it were innate, and unless that form could pave the way for new forms, it would appear a miracle that trust ever occurs. (Baier 1994: p. 107)

Baier counts infant trust as a genuine case of trust even on her restrictive account. She argues that her account of trust does not require advanced concepts or abilities beyond what young children already possess: “One constraint on an account of trust which postulates infant trust as its essential seed is that it not make essential to trusting the use of concepts or abilities which a child cannot be reasonably be believed to possess” (1994: p. 107). We might ask whether Baier’s own account, on which competence and goodwill are ascribed when trusting, satisfies this constraint. Since accidental failures are linked with incompetence, and purposeful failures with ill-will, perhaps the most important sign that one possesses concepts of competence

and goodwill is that one is able to distinguish accidental from purposeful failures. It appears that the ability to distinguish these two kinds of failures is acquired in the second or third year of life (see, e.g. Olineck and Poulin-Dubois 2005).

Here, then, is a way that at least some restrictive views of trust *can* help explain the emergence of practices and institutions: by identifying a form of trust that exists in individuals in an early, perhaps simple form in a way that is *psychologically* prior to those practices and institutions. This is particularly useful in explaining the emergence of stable patterns or practices of cooperation. Furthermore, although innate trust in one's kin cannot be explained as the outcome of central concerns or interests of the relevant actors in an agential sense (assuming the infant is not an agent), it can be so explained in an evolutionary sense.[9] Baier's account does well on this criterion, although this is not a unique feature of her view: the unrestrictive view, and several other restrictive views, will also count infant trust as genuine trust.

In order to develop a positive argument for Baier's view, we need to show that the conception of trust as optimism about the goodwill and competence of the trusted draws a distinction that has important explanatory value in its own right. There is a way of trying to do this. Some psychologists and sociologists draw a distinction between a disposition based on the predictive expectation of reliability, which they sometimes call "confidence," and a trusting disposition, which is not a prediction, but rather an evaluation of the social situation and the person on whom one is in a position to rely. The reason for this distinction is not primarily that it is an intuitive distinction, but that it tracks reliance based on two different kinds of information

[9] There are limits to this. *Failures* of infant trust do not do much to explain the design or evolution of norms and institutions that bolster trustworthiness, in the way that failures of "adult" trust prior to such norms and institutions can help explain their emergence.

processed through different psychological channels, forming two measurably different components of social cognition. As Midden and Huijts explain the idea, "trust becomes the basis of decisions at the point when other assurances are not sufficiently available and (experience-based) confidence is lacking" (2009: p. 744).

From this perspective, what is characteristic of the Trust Game, and indeed of many of the situations where we want to explain the emergence of cooperative practices and institutions in terms of trust, is that one does not have a track record of interaction that gives one access to experience-based information about the prior performance of the other party (or at least not a track record that is applicable in this new situation). Of course we can imagine variations of the Trust Game where we do have such information, but these will be somewhat uncharacteristic for cases where cooperative practices and institutions emerge. What we want to know for explanatory purposes is how people deal with reliance on the other in the *absence* of any kind of experience-based, mechanical, or causal certainty—that is, confidence—about how they will act. This suggests that the Explanatory Constraint may be compatible with drawing a conceptual distinction between trust and confidence. Restrictive accounts of trust can satisfy the Explanatory Constraint insofar as they (a) capture this psychological distinction between trust and confidence, (b) are not otherwise too restrictive about what counts as trust, and (c) suggest a distinctive mechanism by which trust brings about cooperative practices and institutions.

The aspect of Baier's account that best captures this distinction is the idea of *optimism*, a kind of affective slant to how one perceives opportunities in a situation. Karen Jones links this with the distinction just introduced:

> We can be justified in trusting even when we would not be justified in predicting a favourable action on the part of the one trusted. Our

> evidence for trusting need not be as great as the evidence required for a corresponding justified prediction. In this respect trusting is more like hoping than like predicting. (1996: p. 15, quoted in McGeer 2006: p. 241)

This fits with the idea that we have a distinctive ground for our other-regarding attitudes in situations where the evidence of their reliability is not conclusive.

However, without providing a further basis for such reliance, this story is unsatisfactory for two main reasons. First, it remains unclear *why* there does not need to be as much evidence for trust as for a prediction. Even if it is plausible that one is entitled on rational or moral grounds to trust others, because doing so is a matter of respect or decency (as in Ross 1986), it is hard to see this alone as an adequate epistemic or practical reason for doing so. Second, its explanatory value is limited. Optimism is like a wind that blows in the direction of reliance on others. When this wind is blowing, as it were, it helps explain why people tend to rely on others more than the evidence may strictly seem to warrant. However, there are other motivations, more closely connected with warrant, justification, and practical reason, that bring about reliance even when the wind is still. These should be part of the account of trust, according to the Explanatory Constraint. In light of these problems, I will now turn to a view that articulates the distinctive reasons associated with trust: what I will call the "dependence-responsiveness" view. This view can be seen, in a way, as grounding the optimism of trust in a special kind of epistemic and practical reason, and thereby responding to these two objections.

4. Dependence-Responsiveness Reconsidered

Several philosophers have endorsed as central to trust the idea that the person trusted will be responsive to the act of reliance (Pettit 1995; McGeer 2006; Faulkner

2007, 2014). I call this aspect of trust "dependence-responsiveness" (Nickel 2012).[10] It can be conceived of as a necessary component of the trust attitude itself, albeit implicit. Suppose Britta lists Leif as a reference on her job application. In doing so she counts on him to be responsive to her dependence on him, so that if the hiring committee contacts him he will provide a suitable recommendation. This responsiveness to reliance is supposed to lie at the heart of trust.

One argument for the centrality of dependence-responsiveness to trust is that it explains the distinctive wrong of betrayal. Betrayal occurs when a person signals that he will be suitably responsive to one's reliance, and then fails to respond when the need arises. If Leif does nothing to disabuse Britta of the idea that he will provide her with a timely and supportive reference, and then (intentionally) fails to do so, this is an instance of a distinctively manipulative wrong. On the assumption that betrayal is the distinctive wrong associated with broken trust, this appears to support the link between trust and dependence-responsiveness (Faulkner 2007).

Elsewhere I have criticized this view of trust on intuitive grounds, arguing that it is not a necessary condition for trust (Nickel 2012). Although I still maintain this objection, here I would like to argue that the dependence-responsiveness view does a good job satisfying the Explanatory Constraint. As argued in the previous section, restrictive accounts of trust can satisfy the Explanatory Constraint insofar as they (a) capture the psychological distinction between trust and confidence, (b) are not overly restrictive about what counts as trust, and (c) suggest a distinctive mechanism by

[10] Pettit (1995: p. 203) calls this "trust-responsiveness" but I prefer "dependence-responsiveness" for several reasons: the responsiveness doesn't seem to distinguish between reliance and trust, nor does the ascription of it to the person on whom one relies. Also, if the point is to define and rationalize trust in terms of this feature, it seems better not to use trust to define the feature.

which trust brings about cooperative practices and institutions. The dependence-responsiveness view, at least when interpreted broadly, does admirably on all three criteria.

First, it captures the psychological distinction between trust and confidence. The trust-confidence distinction, broadly speaking, is the distinction between a disposition towards reliance based on reliable information that the other will behave in a certain way, or has a certain likelihood of behaving in a certain way, and a disposition toward reliance where this solid information is lacking, so that one cannot strictly estimate the reliability of the other's performance but must place oneself in their hands on some other basis. Confidence is not essentially an attitude that is directed toward persons. It could hold of institutions or even physical objects. But it can hold of persons, for example when an institution is set up such that if one particular employee does not carry out certain tasks for those who rely on them, an appeal can be made to another employee (a supervisor) who will do so. The function of the institution provides reliable guidance to one's expectations, hence this kind of reliance on another is confidence rather than trust. This leaves unspecified what kinds of grounds might count as evidence for trust, which is where the dependence-responsiveness account is useful. Trust, on this view, is distinguished from confidence by the fact that one's placing oneself in the hands of another is expected to make the other more reliable than she would otherwise be. This is clearly not the case in purely bureaucratic institutions. This makes it clearer what the distinctive ground of trust is, as against other kinds of reliance on persons.

Second, this distinctive ground for trust, at least when taken broadly enough, is not too restrictive to be plausibly present in most one-on-one trust relationships and exchanges. It merely requires that the one relied upon will be responsive to the fact

of reliance and will adjust her behaviour to better meet the expectations of the one who relies. It is even plausible that something like this is present in the La Salle–Illinois exchange. The fact that in this exchange it would be appropriate for either side to feel, in a certain sense, *betrayed* if the other side suddenly attacked, or poisoned the food being served, is some indication of this.[11] Some expectation of dependence-responsiveness is compatible with low levels of overall assurance or confidence about the other party. This means that the dependence-responsiveness account can better satisfy the Explanatory Constraint, because it holds of attitudes and interactions prior to cooperative practices and institutions, and can therefore do something to explain the emergence of those practices and institutions.

Finally, the dependence-responsiveness account links with a naturalistically well-grounded mechanism by which cooperative practices and institutions emerge. The mechanism is particularly strong when the idea of dependence-responsiveness is coupled with a means of *signalling* that one is trustworthy. La Salle and the Illinois each show the calumet to the other as a means of inviting reliance, signalling that if relied upon not to do violence to the other, they will behave peacefully. Philosophers with a knack for modelling have used such signalling to explain the success or failure of cooperative interaction in multi-agent simulations (Skyrms 2010). This explanation can be extended if we include the idea of somebody's having a visible "reputation" that is affected by how well they perform. A person who knows that his reputation will be affected by his performance, such as Leif in relation to Britta, will tend to be more reliable (Pettit 1995). This is then an extrinsic reason for being dependence-responsive, which is used in the explanation of the emergence or non-emergence of

[11] It is interesting to consider whether such a feeling of betrayal must imply moral blame.

cooperation. Signals and reputational information then figure as important ways of filling in what the distinctive kind of information consists of, on which trust is based.

In order to have these advantages, however, the theory must be interpreted broadly enough to count various reasons or motivations, including such extrinsic motivations, as instances of dependence-responsiveness. Possible dependence-responsive motivations should include worrying about the possibility of sanctions, trying to preserve one's relationship, and perhaps even strategic motives such as the cunning one that Francesca attributes to George when she plays the Trust Game. Some proponents of the theory seem to distance themselves from this broad interpretation of dependence-responsiveness. In a recent paper, Faulkner argues that it is a mistake to see trust as depending on the kinds of reasons specified within the framework of purely instrumental (what he calls "Humean") rationality (Faulkner 2014). The worry is that this rules out strategic motivations. Faulkner argues that when we trust, we do not paradigmatically think of the performance of the trusted person as a means to the achievement of something that we want, such as the continuation of our relationship with them. At one point, he indicates that in the norm, the trusted person's reason for performance must match the expectations of the person who trusts: "A trusted party S is trustworthy, in a circumstance defined by A's (affectively) trusting S to φ, if and only if S sees A's depending on his, S's, φ-ing as a reason to φ and φs for this reason" (Faulkner 2014: p. 1982). In order to reap the explanatory advantages of the dependence-responsiveness account, we must either take seriously that this further specification (like the rational choice norm in Coleman's account) is not intended to settle the question of what counts as trust; and thereby allow that in some cases the trusting person simply leaves it open to the kinds of reasons to which she expects the trusted person to be responsive.

Some other views of trust can meet the Explanatory Constraint in a similar way. The theory of trust I favour is one on which a person who trusts *normatively expects* the trusted party to perform a certain way, where this does not require dependence-responsiveness in every case. The view differs subtly from dependence-responsiveness in that the norm can—and must, sometimes—be applied even in circumstances where it is not expected to have a knock-on effect (Nickel 2012). Correspondingly, the view does not require that the trusted person responds to the very norm of behaviour on which the trustor's normative expectation is based. However, it allows for that kind of dependence-responsiveness as a normal case. This has explanatory potential insofar as the idea of ascribing and applying norms to others is a fundamental mechanism among humans for changing the balance of reasons and improving compliance with expectations in cooperative exchanges; and also because when these normative expectations fail, we sometimes take other steps to encourage better compliance in the future, such as the establishment of sanctions and institutions.

5. The Concept of Trust: Natural, Social, or Political?

The conclusion of the preceding sections is that some (less) restrictive theories of trust are favoured by the Explanatory Constraint on non-intuitive grounds. What does this show? At best it shows that a theory of trust is better if it allows for explanatory aims, not that this determines the referent of the "folk" concept of trust, which is presumably also the one that philosophers investigate via intuitions, phenomenology, and/or rational reconstruction. Additional argument is needed to link the explanatory aims with the concept of trust with which philosophers have been concerned, showing that this concept, contrary to appearances, is settled by the Explanatory

Constraint rather than by these other methods. In this section I consider a possible argument from scientific essentialism. I claim that this argument is not decisive, and that the explanatory advantages of a less restrictive account of trust must be balanced against other factors.

A preliminary move in trying to establish that scientific aims might determine the concept of trust is to relativize philosophical intuitions. The folk concept of trust is messy. It is common enough to speak of domesticated animals trusting and being trusted, and also to speak of people trusting computer systems, elements of the built environment, and even natural phenomena like the weather. Philosophers often respond to this fact by drawing an intuitive distinction between trust and "mere reliance," or between two kinds of trust (Faulkner 2014). However, without further argument of the kinds we have offered in the previous sections, the appeal to intuition begs the question against the unrestrictive view. To emphasize this point, it is useful to point out that the folk concept, and in particular the distinction between trust and reliance, do not translate identically into different natural languages. Some languages do not draw as sharp a distinction.[12] Our intuitive sense of what genuine trust is, is perhaps not shared universally.

This prepares the way for a further argument for the claim that our concept is determined pragmatically by our scientific explanatory aims, rather than by our intuitions. According to anti-individualism about thoughts, the identity of thoughts embedding certain concepts is not fully determined by that to which the thinker has

[12] In Dutch, "I rely on x" and "I trust x" can both be naturally translated as *ik vertrouw [in, op] x*, and "reliability" and "trustworthiness" can both be naturally translated as *betrouwbaarheid*. In French, the respective pairs are both translated as *faire confiance en* and *fiable*. Of course a distinction can be drawn, but it lies further from the surface than in English.

subjective access. Instead, which concept one is thinking about is partly determined by external facts to which the thinker might have imperfect access: either facts about *scientific essences* or natural kinds which the concepts pick out, or facts about the *social determination* of the concepts (Burge 2007). For example, a person thinking about arthritis is thinking about a rheumatic disease of the joints. Even if her subjective grasp is insufficient for her to have an opinion about whether arthritis can occur in the thigh, she can nonetheless have thoughts that embed *this* concept of arthritis, according to which its occurrence in the thigh is excluded. There are two grounds for explaining this fact: an essentialist ground and a social ground. Arthritis may be a natural kind with an essence, or it may have been settled by the community of scientists who work on arthritis that arthritis only concerns the joints (although perhaps it might have been settled differently). In the case of arthritis, it is not obvious which factor plays the decisive role: is it that medicine has been given the task of defining medical concepts, or the fact that joint disease is a natural kind?

One might think that philosophical concepts are not susceptible to anti-individualist argument. But in fact we find both the essentialist and the social variant of the argument applied by some philosophers to the concept of knowledge. Hilary Kornblith (2014) argues for the claim that knowledge is a natural kind, and that intuition-based investigations of the concept of knowledge are poor guides to the concept. According to Kornblith, the idea of a representational state that correctly corresponds to some external state of affairs has such great explanatory value in describing animal behaviour, and evolutionary success, that it is an essential biological concept:

> When we seek to explain the presence of certain cognitive capacities in a species . . . knowledge enters the picture. The environment makes

> certain informational demands on a species, and cognitive capacities answer to those demands. . . . if we want to know why some individual has a certain cognitive capacity, we will need to advert to the evolutionary explanation . . . The category of beliefs for which these capacities were selected—reliably produced beliefs that are also true—is thus important to ethologists. As I see it, the explanatory importance of the category and its theoretical unity provide reasons for viewing it as a natural kind. (Kornblith 2014: pp. 176–7)

The most straightforward version of an anti-individualist argument applied to trust would claim that trust is a natural kind. Our thoughts about it are externally determined in the same way that our thoughts about water are externally determined by the substance H_2O. In the case of water, even before modern chemistry described the substance, H_2O was what we were thinking about. Those who have theories of water based on intuitions about their experiences of water, perhaps even denying the existence of H_2O, are nonetheless thinking about H_2O.

Is there a case to be made that trust is a natural kind like this, and that intuitions are a poor guide to the concept? Philosophers since Aristotle have made serious comparisons between the cooperative, social behaviours of humans and other animals. The social disposition of humans to rely on other members of their species might be considered a natural kind, particularly if we understand this disposition in the simplest possible way, perhaps in the style of Baier's "infant trust," so that other social animals partake of it. Baier makes the point that infant trust does not even require that the child has a fully developed capacity of choice. The ability to distinguish intentional from accidental non-reliability is also shared to a degree by non-human animals. What is distinctive about human trust, on this view, is that unlike

other animals we are also able to engineer the environment by creating complex signals, social norms, and institutions that change the balance of motivation for those relied upon in situations of uncertainty, rendering social interaction within societies more stable and reliable.

However, it appears from the argument of the previous sections that a cognitively richer view of trust may also offer a compelling explanation of cooperative practices and institutions as a social phenomenon. The dependence-responsiveness view requires that the trustor attribute a capacity of choice or decision to the one trusted. On this view, the trustor expects the trusted to be aware of the trustor's reliance and, in virtue of this, to become more responsive in fulfilling their expectations. This is a complex attribution of intention and choice, requiring a complex cognitive apparatus not likely to be shared by many other animals. Although such a trust disposition might also conceivably be a natural kind, the style of argument that Kornblith deploys to show this in the case of knowledge would not apply. His argument turns on the claim that knowledge is a category that we need at many points to explain evolutionary success across many species. If trust as dependence-responsiveness is a natural kind, it is anthropogenic.[13]

Burge's original argument for anti-individualism emphasizes the idea that the concept embedded in our thoughts is a matter of use within our language community, the "social environment," rather than Kornblith's idea that natural kinds

[13] Kornblith appears to limit himself to biological kinds when arguing that knowledge is a natural kind. When criticizing Edward Craig's (1990) view that knowledge is a concept that humans (including early humans) needed in order to identify good sources of information, he never considers the possibility that an anthropogenic kind could also be a natural kind (Kornblith 2014: Ch. 12).

directly fix the contents of our thoughts.[14] At least in some cases, it is because of the authority socially accorded to scientists that scientific explanatory values determine which concept is embedded in our thoughts. This point does not require that there be extralinguistic, metaphysical truths involving biological or other kinds. It is only required that the actual concept of arthritis, for example, is fixed the way it is because scientists have found this concept useful, and because the broader linguistic community gives them authority over the concept. The fact that the authority traces back to the community does not imply that an individual can deploy an idiosyncratic, non-standard concept, any more than the fact that laws trace their authority back to the community implies that a person can opt for an alternative law.

As applied to trust, however, there are two serious problems with this idea. The first is that there is little interdisciplinary agreement about trust. That medical scientists' understanding of arthritis should determine the boundaries of the concept is plausible, but it is not similarly plausible that biologists, rather than psychologists, political scientists, economists, or philosophers, should determine the boundaries of the concept of trust. The second problem is that trust is a concept with heavy social and political significance in modern times (see, e.g. Baberowski 2014). Concepts like this, such as *legitimacy* and *democracy*, are public property, maybe even contested territory. The boundaries of such concepts are not dictated by experts alone. It appears, then, that there is no decisive reason to regard the concept of trust as being determined by a natural kind, nor by social facts that give experts authority

[14] In a postscript to "Individualism and the Mental," Burge stresses that the argument in the original paper concerns the role of the social environment in fixing our concepts, but states that on his current view both the social and the physical environment play a role in fixing the concept (Burge 2007 p. 152).

over the concept. Perhaps philosophers even have a special role in mediating public and expert discourse about trust, giving *them* a kind of special interpretive authority over the concept. This suggests that the argument from explanatory potential is not a knock-down argument, but that it must be taken into account alongside other considerations.

6. Conclusion

Philosophical accounts of trust based on intuitions and the phenomenology of trust have strong prima facie appeal, and the argument from explanatory potential does not bypass our ordinary intuitions in the way that an analogous argument might in the case of the concept of arthritis or H_2O. My conclusion is pragmatic: philosophers should take the argument from explanatory potential on board as a way of grading accounts of trust. Other things equal, it is better to have minimal restrictions on motives, to interpret these restrictions psychologically in such a way that permits them to be ascribed broadly to many agents in many situations, and to mark genuinely explanatory distinctions with the concept, identifying features that make a difference to cooperative behaviour and institutions. I have argued here that the argument from explanatory potential may favour a view of trust on which it is strongly linked with dependence-responsiveness: the feature that when one person trusts another person, the first person assumes that the second person, aware of the reliance of the first, will (be more likely to) choose to be reliable. In addition, I have argued that this notion of dependence-responsiveness should be interpreted broadly

to include responsiveness to reputational considerations and other strategic motives.[15]

References

Baberowski, J., ed., 2014. *Was ist Vertrauen? Ein interdisziplinäres Gespräch*. Frankfurt: Campus.

Baier, A., 1994. *Moral Prejudices: Essays on Ethics*. Cambridge, MA: Harvard University Press.

Berg, J., J. Dickhaut, and K. McCabe, 1995. Trust, Reciprocity, and Social-History. *Games and Economic Behavior* 10(1): pp. 122–42.

Buchak, L., 2016. Decision Theory. In A. Hájek and C. Hitchcock, eds., *Oxford Handbook of Probability and Philosophy*. New York: Oxford University Press.

Burge, T., 2007. *Foundations of Mind: Philosophical Essays, Volume 2*. New York: Oxford University Press.

Camerer, C. and K. Weigelt, 1988. Experimental Tests of a Sequential Reputation Model. *Econometrica* 56(1): pp. 1–36.

Coleman, J. S., 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Cook, K. S., R. Hardin, and M. Levi, 2005. *Cooperation without Trust*. New York: Russell Sage Foundation.

Craig, E., 1990. *Knowledge and the State of Nature*. New York: Oxford University Press.

[15] Thanks to the editors of this volume, and to participants at an Oxford workshop in 2014, for comments on earlier versions of this paper. This research has been funded in part by a grant from the Socially Responsible Innovation (MVI) programme of the NWO, for the project "Medical Trust Beyond Clinical Walls."

Faulkner, P., 2007. On Telling and Trusting. *Mind* 116: pp. 875–902.

Faulkner, P., 2014. The Practical Rationality of Trust. *Synthese* 191: pp. 1975–89.

Hardin, R., 2006. *Trust*. Cambridge: Polity Press.

Hieronymi, P., 2008. The Reasons of Trust. *Australasian Journal of Philosophy*, 86: pp. 213–36.

Hobbes, T., 1968 [1651]. *Leviathan*. Ed. C. B. Macpherson. New York: Penguin Books.

Hume, D., 1998 [1751]. *An Enquiry Concerning the Principles of Morals*. Ed. T. L. Beauchamp. New York: Oxford University Press.

Johnson, N. D. A. A. Mislin, 2011. Trust Games: A Meta-Analysis. *Journal of Economic Psychology* 32(5): pp. 865–89.

Jones, K., 1996. Trust as an Affective Attitude. *Ethics* 107: pp. 4–25.

Kornblith, H., 2014. *A Naturalistic Epistemology: Selected Papers*. New York: Oxford University Press.

Kreps, D. M., 1990. Corporate Culture and Economic Theory. In J. E. Alt and K. A. Shepsle, eds. *Perspectives on Positive Political Economy*. Cambridge: Cambridge University Press: pp. 90–143.

McGeer, V., 2006. Trust, Hope and Empowerment. *Australian Journal of Philosophy* 86(2): pp. 237–54.

McLeod, C., 2002. *Self-Trust and Reproductive Autonomy*. Cambridge, MA: MIT Press.

Midden, C. J. H. and N. M. A. Huijts, 2009. The Role of Trust in the Affective Evaluation of Novel Risks: The Case of CO_2 Storage. *Risk Analysis* 29(5): pp. 743–51.

Nickel, P., 2007. Trust and Obligation-Ascription. *Ethical Theory and Moral Practice* 10: pp. 309–19.

Nickel, P., 2009. Trust, Staking, and Expectations. *Journal for the Theory of Social Behaviour* 39: pp. 345–62.

Nickel, P., 2012. Trust and Testimony. *Pacific Philosophical Quarterly* 93(3): pp. 301–16.

Olineck, K.M., and D. Poulin-Dubois, 2005. Infants' Ability to Distinguish Between Intentional and Accidental Actions and Its Relation to Internal State Language. *Infancy* 8(1): pp. 91–100.

Parkman, F., 1983 [1878]. La Salle and the Discovery of the Great West. In Francis Parkman: France and England in North America: volume I. Ed, D. Levin. New York: Library of America: pp. 713–1054.

Pettit, P., 1995. The Cunning of Trust. *Philosophy and Public Affairs* 24: pp. 202–25.

Ross, A., 1986. Why Do We Believe What We Are Told? *Ratio* 1: pp. 69–88.

Simpson, T., 2012. What is Trust? *Pacific Philosophical Quarterly* 93(4): pp. 551–69.

Skyrms, B., 2010. *Signals: Evolution, Learning and Information*. New York: Oxford University Press.

Utz, S., U. Matzat, and C. Snijders, 2009. On-Line Reputation Systems: The Effects of Feedback Comments and Reactions on Building and Rebuilding Trust in On-Line Auctions. *International Journal of Electronic Commerce* 13(3): pp. 95–118.

Walker, M.U., 2006. *Moral Repair: Reconstructing Moral Relations After Wrongdoing.* Cambridge: Cambridge University Press.