Trust in Technological Systems

Philip J. Nickel

Draft Version 2011-03-22

To appear in M.J. de Vries, S.O. Hansson, and A.W.M. Meijers, eds., *Norms and the artificial: moral and non-moral norms in technology* (Springer).

I. Introduction

Technology is a practically indispensible means for satisfying one's basic interests in all central areas of human life including nutrition, habitation, health care, entertainment, transportation, and social interaction.[1] It is impossible for any one person, even a well-trained scientist or engineer, to know enough about how technology works in these different areas to make a calculated choice about whether to rely on the vast majority of the technologies (s)he in fact relies upon. Yet there are substantial risks, uncertainties and unforeseen practical consequences associated with the use of technological artifacts and systems. The salience of technological failure (both catastrophic and mundane), as well as technology's sometimes unforeseeable influence on our behavior, makes it relevant to wonder whether we are really justified as individuals in our practical reliance on technology. Of course, even if we are not justified, we might nonetheless continue in our technological reliance, since the alternatives might not be attractive or feasible. In this chapter I argue that a conception of trust in technological artifacts and systems is plausible and helps us understand what is at stake philosophically in our reliance on technology. Such an account also helps us understand the relationship between trust and technological risk, and the ethical obligations of those who design, manufacture and deploy technological artifacts.

[1] We may distinguish between welfare interests and ulterior interests. Welfare interests are those interests which are assumed to be had by any person, such as freedom of movement, ownership of property, life, health, etc. Ulterior interests are related the specific goals of individuals, such as the desire to own a motorboat, to travel to Hawaii, to start an organic farm, etc. See Feinberg (1984).

First, a terminological remark. I will use the terms *artifact* and *technological system* in a non-standard way in this chapter. By an artifact, I mean a relatively discrete object or component part of an object that has been created intentionally for a particular function. By a technological system, I mean a constellation of artifacts that work more or less in conjunction with one another and that create or enable some form of activity. An example of an artifact is a washing machine, or a heating element in a washing machine. An example of a technological system is an array of appliances and component parts, appliance repair equipment, electrical and water supply lines, drainpipes, electrical outlets, detergents, clothing instruction tags and so forth that together make our activity of washing clothes possible. This contrasts with the standard engineering use of the term *system* as a set of parts working together for a given function. In order to emphasize my non-standard use, I will use *technological system* instead of *system* by itself.

I begin, in this and the next section, with an explanation and justification of the very idea of trust in technology. It is not an idea widely accepted among mainstream philosophers of trust, despite recent interest in the topic among theorists of technology.[2] Philosophers often distinguish between judgments of reliability on the one hand and genuine trust on the other, pointing out essential differences between these two attitudes.[3] Whereas a judgment of reliability consists of a purely predictive expectation of performance, trust consists of a normative or moral expectation of performance, perhaps associated with characteristic moral "reactive attitudes" such as betrayal, blame and resentment. Technological artifacts are mentioned as paradigmatic examples of things about which we make judgments of reliability rather than things we can genuinely trust: "In cases where we trust and are let down, we do not just feel disappointed, as we would if a machine let us down. We feel betrayed" (Holton

[2] Recent special issues of *Knowledge, Technology and Policy* and *Ethics and Information Technology* are devoted to trust in technology (Taddeo 2010).

[3] People often contrast trust with mere reliance,but this is not a suitable comparison. Reliance is way of acting, whereas trust is an attitude. For our comparison we must therefore find an attitude associated with "mere" acts of reliance — here I call this a *judgment of reliability*.

1994, 66). Another prominent philosopher of trust writes: "Trusting is not an attitude that we can adopt toward machinery. I can rely on my computer not to destroy important documents or on my old car to get me from A to B, but my old car is reliable rather than trustworthy. One can only trust things that have wills…" (Jones 1996, 14). On this view, although we sometimes use the word 'trust' to describe our reliance on technology, this is just a relaxed manner of speaking. Genuine trust is an evaluatively rich interpersonal attitude concerning the motives of the trusted person, but reliance is different. It is as if in one's reliance on technology, one couples a judgment of the means available for achieving one's ends with the engineers' definition of reliability: "The probability that an item will perform a required function without failure under stated conditions for a stated period of time" (O'Connor et al. 2002, 2). It seems that the burden of proof falls on one who wants to show that there is an interesting sense in which we can trust technological artifacts going beyond this conception (see Nickel et al. 2010).

There is one obvious strategy available for showing that trust in technology is something more than a mere reliability judgment. It has sometimes been suggested that genuine trust in an artifact is possible so long as one thinks of it as having as its implicit object the people and institutions responsible for the creation and maintenance of the artifact. For example, when I trust a bridge not to crumble as I cross it, the real object of my trust is those who built the bridge and those who are administratively responsible for its maintenance (Origgi 2008, 12). In this way we do not have to countenance artifacts as free-standing objects of trust, because we can explain trust in artifacts as a kind of trust in persons. Trust in a bridge reduces to or is fully explained in terms of trust in those who are responsible for creating and maintaining the bridge. My trust that my Honda will start is to be explained in terms of my implicit attitude toward the Honda Motor Company and its employees.

This strategy cannot fully explain the idea of trust in artifacts, however. For when I turn the ignition of my car, I rely on the firing of the ignition system, not the creation of a car that has an ignition system that fires. The firing of the ignition system is not an action performed by the employees of the Honda Motor Company, nor would we normally say that these employees caused my car to start. It is instead an event caused by the artifact they created, and it is this event that I rely upon. The following principle seems evident:

(T→R) If I trust some entity E to perform in a particular way, then it must be a salient possibility that I (or somebody I care about) could rely on E to perform in that way.

Since I do not rely on the employees of the Honda Motor Company to cause the ignition system to fire, it follows that I do not trust them to do so. The entity I trust for this performance must be the car itself. To understand trust in technology we must therefore look at the question of whether I can trust an artifact itself.

I will argue for an account of trust, the Entitlement Account, that makes sense of direct trust in artifacts. The basic idea is that in trusting, we exhibit a willingness to rely for something we value on the performance φ of some entity (such as a person, artifact, or system) about which we are committed to saying both that (a) it is worth relying on φ; and (b) we are *entitled* to rely on φ. The attitude of entitlement is a normative expectation about the performance of the entity, to be explicated in the next section.

This account explicates a core attitude common to several kinds of trust including institutional trust, self-trust (both mental and bodily), and trust in artifacts. Although applying this account in detail to self-trust and institutional trust would lead us too far away from the purpose of this chapter, a brief example is useful to indicate the broader theme. Consider how I trust my memory. Normally, when I make plans today, I trust my own future memory faculties to guide my planned action. If my memory fails, I am likely to feel frustrated or disappointed in a way that reveals an attitude of entitlement to the performance

of my own faculties. This feeling of frustration goes beyond what could be explained by a mere judgment of reliability (summed up in the idea that it is *worth it* to rely on my memory), although this is also a component of self-trust. It exhibits a normative dimension. In the next section I will develop an argument along these lines concerning trust in technology.

II. Technological Trust and Entitlement to Performance

The first element of trust in a technology T is that T is likely enough to perform in some desired way φ, compared with other alternatives, that it is worth staking something of value on φ. Without some such attitude, it is hard to explain why a person is disposed to act in a way that presupposes or counts on the performance. But this first element by itself is not sufficient for trust, for in trust one also has a normative attitude toward the entity trusted. Thus we need a second element in any account of trust in technology, an attitude of entitlement to the technology's performance. In this section I wish to make it plausible that such an attitude exists and that we can make some sense of it folk-psychologically. I do not aim here to show that it is legitimate in the deepest sense. Plausible attitude-ascription does not depend upon the ultimate legitimacy of those attitudes. For example, philosophers who argue that attitudes of blame are illegitimate on the ground that there is no such thing as freedom of the will, and hence that there is no suitable object of blame, do not thereby infer that blame does not exist.

To show that an attitude exists, the central step is to show that it explains a cluster of phenomenological and psychological observations by providing a rationale for those observations, and can be given a plausible description in itself. In the case of our reliance on technology, there are some ordinary phenomena that cannot easily be explained without supposing the existence of an artifact-directed entitlement-attitude. This becomes apparent when we focus on negative attitudes toward technological failure such as anger,

disappointment, and frustration. All of these negative attitudes can take an artifact or technological system as their object. An academic survey article on anger's role in human development states that it " is associated with infants' attempts to master the physical environment, and … elicits behavioral strategies from infants that serve regulatory functions and contribute to problem solving" (Lemerise and Dodge 2000, 596). Technological artifacts and systems form a large part of the physical environment throughout human development. This is why a car that won't start or a computer that won't print the last chapter of one's dissertation can elicit anger.[4] Frustration and disappointment at the artifact are also common (as indicated by Holton in the passage quoted above, *op cit.*). These attitudes are distinct from (though compatible with) disappointment in an outcome, anger at oneself, and anger at those who created an artifact.

What is the rationale for such attitudes? If the only attitude we ever took toward an artifact or technological system were that it is sufficiently likely to perform in a certain way that it makes sense to rely upon that performance, there would be no rationale for artifact-directed disappointment or anger. For first of all, it is logically consistent with this judgment that the artifact or system does not in fact perform. The judgment at the core of my original attitude has not been straightforwardly contradicted by later events, since likelihood and non-occurrence are not strictly opposed. At most, if an event judged likely does not occur this requires some revision of one's later judgements of likelihood, and an adjustment of one's plans. In itself, it seems implausible that a "cold" epistemological attitude such as prediction would generate object-directed disappointment or even anger when it failed. If anything, its failure should generate other epistemic or intellectual attitudes such as curiosity or puzzlement. Therefore, in order to explain artifact-directed anger and disappointment, it is useful to suppose that people sometimes have a richer normative attitude toward artifacts

[4] Or even aggression: destruction of objects is a clinical criterion for hostile aggression, although this is often a surrogate for interpersonal aggression (Ramírez and Andreu 2006).

when they rely on them. I propose an attitude of *entitlement*, with the artifact's performance as its object.

This key notion, entitlement, has two established philosophical meanings. In one sense, it refers to the positive normative status of an attitudinal state, such that one cannot, other things equal, be criticized for having that attitudinal state. For example, some epistemologists think we are *entitled* to the belief that physical objects exist, even though they also hold that we have no ultimate *justification* for this belief since there are no sound arguments conclusively establishing its truth. We cannot be criticized for holding the belief, for it provides support to many other beliefs that are themselves confirmed by experience, consistent with further beliefs we possess, and useful or even indispensible to our lives (Wright 2004).

Second, entitlement is sometimes understood as a *right* to goods or services from another person. For example, I might be entitled to three apples belonging to another person in virtue of having won a bet with that person. In this sense an entitlement is a Hohfeldian "claim right," consisting of a liberty to receive these apples together with an obligation the other person has to provide me with them (Wenar 2008). In the case where I come across apples not belonging to anybody, my claim right to pick them is a negative right: it consists of a liberty to take them, together with an obligation that others not interfere with my doing so.

It is sometimes assumed that the idea of an entitlement to an artifact's performance must fall into either the epistemic type or the rights type: entitlement is either the warrant for a *belief* about what I can reasonably predict of the artifact, or it is a claim right implicitly directed toward a person or persons responsible for ensuring that it performs a certain way (Franssen 2009). For example, my entitlement to my car's starting is understood either as an epistemic prediction about the car's starting, or as a claim right according to which "I can

hold the manufacturer, or, in the case of a second-hand car, the car dealer, to his or her part of a deal we made" (*ibid.*, 941). I wish to deny this. The notion of entitlement I wish to elucidate is identical neither with an epistemic entitlement, nor with a claim right, even though it has something in common with each. With the epistemic notion of entitlement it shares the feature that one counts upon a performance (here both practically and epistemically) the likelihood of which cannot be given a conclusive justification. Although on many given occasions one can question or demand justification for one's reliance on a technological artifact, or refuse to use the artifact because such justification cannot be found, in our civilization it would be impossible to do this in general without foregoing the normal pursuit of one's interests. Wittgenstein (1969) made a case that doubt, as well as demands for justification, can only be pursued in everyday life against a backdrop of confident but unjustified belief. Similarly, in a technological age, doubts about technological reliability can only be pursued in everyday life against a backdrop of normal, confident reliance. This generates a default practical entitlement to reliance, in the absence of specific reasons for doubt.

The notion of entitlement I have in mind also has something in common with a claim right. First of all, it is normative: it supposes that the entity ought to perform a certain way, in an evaluative sense. Such an entitlement could be justified in a number of ways, drawing in the first instance on the fact that the artifact has the function of performing in that way — that is what the artifact is *for*, what it has been created (and probably advertised and sold) to do, or what it evolved to do. This entitlement could then be used in some circumstances to justify further normative, moral and legal claims, although it is not identical with those claims. For example, "Failure to φ should be compensated for by the manufacturer," "Failure to φ is a sign of bad design," "The designer is blameworthy for failure to φ." Some of these further claims might indicate rights-claims, although they need not do so.

This notion of normative, default entitlement is common to many kinds of trust, including artifactual trust, self-trust and institutional trust. Here are some examples.

(A) A baker prepares dough on Tuesday evening, intending to punch it down and roll it out the following afternoon. The baker takes herself to be entitled on that day to rely on her own memory and practical abilities the next day — to complete the action.
(B) Quoting David Lewis: "In my hometown of Oberlin, Ohio, until recently all local telephone calls were cut off without warning after three minutes. Soon after the practice had begun, a convention grew up among Oberlin residents that when a call was cut off the original caller would call back while the called party waited" (2002, 43). Suppose one's call is cut off. In such a situation one takes oneself to be entitled to rely on the original caller to call back.[5]
(C) I work on a manuscript for an entire day and at the end of the day I hit the "save document" button in my word processor, planning to open it again the next day. I take myself to be entitled to the document's latest version being recorded to the hard drive.

A normal way to express this notion of entitlement is to say that some entity (one's memory, one's neighbor, one's computer) *is supposed to* do something. It can also be expressed using *ought* or *should*, where these words are not necessarily taken to express any claim of moral obligation or requirement. I can say, in a normatively rich sense, that I was supposed to remember something or that I ought to have remembered it, without supposing that it is my moral obligation to remember it.

Cases (A)-(C) are not cases in which a person has a claim right to performance, for this would suppose wrongly that there is somebody with a duty to provide the performance in question, or to refrain from interfering with the performance, in each case. This would in turn license a moral emotion such as blame or betrayal if performance failed — a right of moral complaint. Such a moral emotion would not be appropriate in these cases, because there is nobody with such a duty. Furthermore, although belief-states are involved in these cases, the epistemic sense of entitlement is too thin to capture the force of these examples. For, first, the issue is what one can depend on in one's actions, not to what extent one has a particular belief. And second, the normativity associated with epistemic entitlement as I

[5] This entitlement holds even if one is not oneself the person originally called. For example, suppose my brother was phoned but had to leave suddenly just after the call was cut off, and before the conversation had finished. If I am standing nearby, I am entitled to rely on the original caller's calling again.

explained it above is too weak to account for the attitudes one has about performance. For in these cases one is licensed to be irritated, angry or disappointed *at some entity* if performance fails — in the first case above, at oneself; in the second at the caller; and in the third at the computer program. I am not similarly licensed to be angry or disappointed at some entity in the event that some belief in which I am epistemically entitled merely turns out to be false.[6] Furthermore, I am not licensed to be angry or disappointed at just any object for which I stake something of value on its behaving a certain way. For example, I am not licensed to be angry or frustrated at a coin that comes down tails when I have bet heads.[7] My attitude toward the coin in such an instance is different than it would be at a coin that failed to register as proper currency in a vending machine, for example. It is the latter attitude that I wish to describe as an attitude of entitlement, a kind of normative expectation.

I have said that anger or frustration is a normal response to breakdowns of this type of trust. It is often argued, by contrast, that blame and betrayal are the central negative attitudes associated with broken trust (Faulkner 2007; Hieronymi 2008; Holton 1994). But while betrayal and blame have been well-studied in connection with trust, little attention has been paid to normatively-laden emotional responses toward people who act incompetently but non-culpably, toward one's own performance failures, toward institutions that fail, and toward artifacts that fail. In cases of others' non-culpable failures of competence, disappointment can be highly appropriate. In cases of one's own failure to perform as expected (forgetting, failing to notice something, failing to execute bodily movements successfully, etc.), self-directed frustration is very common. Institutions that fail without the known culpable failure of any individual can induce appropriate anger. Such attitudes are pervasive. Similar reactions toward artifacts also seem appropriate, and although they could

[6] An exception is the case in which I am told something false by another person. But in that case I have reason to be angry at the person not just because her claim happened to be false but because she wrongly presented the claim as true.

[7] This point is due to Peter Kroes.

also give rise to feelings of blame and betrayal toward the artifact's designer or manufacturer, they are not identical with those feelings.

Our answer to the question whether technological artifacts can normally be an object of anger, say, partly depends on our ability to *make sense of* an attitude so directed. Charity encourages us to interpret people's attitudes and behaviors in the way that can be given the most rational explication. This may make it difficult to notice the attitude of entitlement that I have placed at the center of trust in technology, because it is easily conflated with a rights claim and with the interpersonal attitudes that accompany that claim. Moral philosophers have been reluctant to admit the possibility that object-directed negative attitudes not conveying a determinate moral complaint can be appropriate. In part, I think, this is simply because they have not paid sufficient attention to the rich array of such attitudes. In part it may be because prevailing moral theories attempt to vindicate emotions of moral complaint, but pay little or no attention to object-directed complaints of other kinds.

III. Trust, Risk and Technological Systems

Let us define risk as the non-trivial (proper) possibility that a serious negative effect will arise in the future, and let us define safety of some process or product as a threshold of low risk associated with the process or product such that it is objectively reasonable to accept the risk.[8] What is the relationship between the following claims about safety and trust?

(A) We have conclusive reason to believe that a technological system S is safe (*subjective certainty* that S is safe).

(B) We trust S.

These two claims are often both true for a given technological system, but neither directly implies the other. On the one hand, even in the case where we think a system like S is a good

[8] These definitions of risk incorporate value notions such as "serious negative effect" and "reasonable to accept." For an argument that the notions of risk and safety are fundamentally normative, see the chapter by Möller in this volume.

idea, and are in a position to rely on S, (A) does not imply (B). We may have conclusive reason to believe that S is safe, but still not trust S. One reason for this is that safety is not the only thing we care about. We may think there are other technologies that perform better than S, or we may think relying on S is not worthwhile given the benefits it produces.

More interesting is the question whether (B) implies (A). Normally it is true that when I trust a technological system I regard it as safe, i.e. objectively reasonable to accept the risks it poses. But the notion of 'objectively reasonable', and thus the notion of safety as I have defined it, seems to be based on a situation in which one has more than one good option. (A system does not become safe just because there are few alternatives to it, after all.) Trust, on the other hand, might seem somewhat more adaptable to situations in which there are few options. In cases where I do not have many good options besides S and S falls short of safety, we might still want to say that I choose willingly to rely on S, and that I have normative expectations about S's performance, and thus that I trust S. We are of two minds about the relation between willingness and good options. In one sense, when I hand my wallet to an armed robber, this is voluntary, since I choose that action over likely injury or death. In another sense, I act involuntarily because I have no other reasonable option.[9]

This difficulty sometimes arises when a new technological system comes into being that completely replaces an old way of pursuing an activity, leaving no easy way to opt out of reliance. For example, a card-based system for gaining access to public transportation might be introduced, or a nuclear power plant might be located near one's long-time place of residence. Suppose I have my doubts about whether these systems are safe (or secure): I feel they may present significant risks to my privacy or life, respectively. Although the question of whether I *feel* trust toward these systems may still be in order, my disposition to rely on the systems is difficult to assess. I may have little choice but to rely on the system, therefore

[9] This issue is famously discussed by Aristotle, who already sensed the difficulty of settling it in book III of his *Nicomachean Ethics* (Aristotle 1998). Aristotle himself calls coerced actions voluntary because their immediate cause lies in the person who acts.

in one sense I am so disposed. But I may remain (somewhat) against doing so, were I to have a fully adequate choice in the matter. In that case, my trust or distrust has no immediate practical point, but merely registers my feelings, gesturing indistinctly toward some hypothetical situation in which I might be in a position to choose.

Some contexts like this, in which people rely on pervasive technological systems and where doing so would be difficult to avoid, have raised prominent questions about trust in the academic literature. Examples are electronic voting systems (Pieters 2006), electronic patient dossiers (Barrows and Clayton 1996), large-scale energy projects (Viklund 2003), and GMO foods (Frewer et al. 2003). For example, a recent paper describes trust attitudes toward a new system for e-voting in the United Arab Emirates (Salem 2007). As described in the paper, in the UAE there is no election law, and voting is carried out by an electoral college of several thousand citizens appointed by the rulers' courts. The electoral college is more than four-fifths male. In this context, an electronic voting system has been introduced including an electronic system for the registration of electors and candidates, a system for verifying the identity of electors, and a vote casting system with kiosks, a voting database, an encryption system, and a vote counting and results-presentation system. Considering the question of trust and how to enhance it through "knowledge management," the author observes that "Perhaps, the clearest indicator of voter trust and satisfaction with the system" is that the electoral authorities received no requests for a recount and that there were no recorded incidents of voter fraud (*ibid.*, 9). The question of how citizens who were not allowed to vote felt about this technological system is not addressed.

Unless citizens wish to leave the UAE, the existing voting system and its technological realization is a fact of life for them, completely embedded in the emerging machinery of democracy there. It is unclear what meaning we should assign to trust in this context. If we ask whether people trust the system, are we asking whether they acquiesce in a

system that poses uncertain risks, that currently violates international norms of democratic political representation, and that they can hardly avoid? Or whether *if* they were free to choose this system or another available alternative, they would actually be disposed to choose it and would feel entitled to a certain (safe, secure, fair, privacy-respecting) performance from the system? The former is a practically salient question, but has little to do with trust, since an affirmative answer is compatible with complacent or even grudging tolerance, and a high perception of risk. The latter is not as practically salient, but asks a question more in line with the willingness we associate with trust, one that has greater ethical relevance to the legitimacy of the voting system, and one that tracks perceptions of safety. It may only be practically relevant to ask this latter question with regard to those instances of reliance on technological systems where one comes close to having a free choice, such as whether to use the Internet (or a particular website) to find health information (e.g., Vedder 2003). It is usually asked against a backdrop of widespread and largely unquestioned reliance on existing technological systems and the artifacts that comprise them, singling out one element for assessment. If this is how we understand questions about trust in technological systems, a conceptual connection between trust and perception of safety is maintained. Trust in a technological system implies imputing a level of safety to that system that one would be willing to accept given a real choice (assuming that the other condition, the normative expectation of a certain performance, is also met).

IV. The Ethics of Trust and Technology

In this final section I discuss the ethics of trust and technology, an area in which ethics and epistemology are closely linked. On the one hand, the ethics of trust requires an ethics of belief. Trust centrally involves doxastic states regarding expected future events. It is partly an estimation that some future event is likely enough to be worth relying upon. But

trust is also practical; it inherently disposes one toward the act of reliance, even if the opportunity to do so is not always present. Because of its practical dimension, the ethics of trust is not a pure ethics of belief (if there is such a thing). A person's decision to rely on a given technological artifact or system is influenced by pragmatic factors, such as what alternatives are available (if any), the time it would take to research these alternatives, the benefits she stands to gain if the artifact or system performs as expected, and so on. These pragmatic factors may also affect what she expects from the technology normatively, what she feels entitled to from the technology. Furthermore, the person who relies on technology is rarely in a perfectly free and epistemologically privileged position. She must rely on technology under conditions of time pressure, insufficient cognitive resources, and without full information.

In my view, a person's epistemological and practical condition is improved when she has adequate, sound justification for her trust, going beyond the mere entitlement discussed in section II. I will call this the *justified trust ethic*. It is a particular application of a plausible minimal rationalist principle which states that in making important decisions it is best to do so on the basis of adequate reasons. On my account of what trust in technology is, one's justification for it consists in adequate reasons to believe (a) that the technology is sufficiently likely to achieve a performance φ that it is worth relying upon it, and (b) that one is in fact entitled to φ. *Having* the justification consists of having access to these reasons, in the sense that one grasps and recognizes the reasons. Having the justification does *not* require being able to articulate, weigh, or support these reasons with more fundamental argument. That would be too intellectualized a conception of having reasons for trust. Even if the notion of a justification is not hyperintellectualized, however, coming to have a justification for one's trust requires scarce cognitive resources such as attention and short-term memory. For this reason, the justified trust ethic must be qualified: coming to have a

justification for one's trust sometimes takes up too much time or demands excessive cognitive resources, and must be balanced against other factors.

There are two serious arguments against the justified trust ethic. The first is based on skepticism about whether most people can actually have an adequate justification for their trust. They stress that most people are incapable of rationally processing relevant information about risk, probability and benefit. For example, people tend to value the elimination of risk more highly than the reduction of risk by an equivalent interval (Tversky and Kahneman 1992, Tversky and Fox 1995). They evaluate risks differently when probabilistically equivalent but notationally different ways of presenting information are used (Carling *et al.* 2009). This and other psychological research suggests that non-experts (and probably experts as well, when they are not thinking formally) cannot recognize the reasons that support their judgments of risk. Indeed, it is commonly argued (or just assumed) that trust is an alternative to making serious judgments about risk, an easier heuristic that allows one to take risks without doing the cognitive work of gathering information oneself and evaluating it.

Perhaps for these reasons, those who write about trust and technology often presuppose that that the designers, manufacturers and deployers of technology have only two trust-related tasks: first, to make the technology as reliable as possible at doing the things it is supposed to do, where this is understood to include safety — thus to make it trustworthy; and second, to persuade possible users that it is sufficiently reliable and safe, thus trustworthy, using whatever psychologically effective means are available. I will call this the *trustworthiness ethic*. In arguing that one should *design for trust*, it is sometimes unclear whether it is the trustworthiness ethic or the justified trust ethic that predominates, since both have as their goal the elicitation of a user or client's trust in some contexts.[10]

[10] There is also a more specialized sense of 'design for trust.' Vermaas *et al.* discuss a special case in which the item to be designed is itself a trust-facilitating information technology (2010).

In addition to the argument that justifying trust is too demanding, a second positive argument that might be offered in favor of the trustworthiness ethic, and against the justified trust ethic, is that relying on trustworthy technology makes people better off, whereas having a further justification for that reliance does not add any additional benefit. This point is a variant of what has been called the "Meno problem" (named after Plato's dialogue *Meno* in which the problem was first formulated) for the value of justification (Plato 1961; Pritchard 2008). The idea is that having a justification for one's true belief does not make one any better off than simply having a true belief. To use Plato's example: one may travel to Larissa if one has a true belief about which road to take; having a justification for that true belief does not get one to Larissa any more effectively.

Elsewhere, I make a case for the justified trust ethic, claiming that in certain contexts it generates an obligation on the part of the designer, manufacturer, or deployer of a technology to provide evidence of trustworthiness to people in a position to trust that technology.[11] In the remainder of this chapter I will not repeat that argument, but instead expand upon it, defending the justified trust ethic against the two objections to it described above. Thus, first, I will argue that there are often kinds of evidence available relevant to the components of trust that can plausibly be recognized and grasped by the person in a position to trust; and second, that a justification acquired by the person, comprised of this evidence, has value for her. This clears the way to hold that those designing, making, or deploying technology to be trusted must both provide sound evidence of trustworthiness, and must not provide unsound or irrelevant evidence that could mislead the potential user. Of course this conclusion is only intended to apply to technologies that are plausibly trustworthy given certain reasonable aims of the potential user; if the technology is not trustworthy and the

[11] Nickel 2011. I will simply say "designer" and "user" in what follows for the sake of economy.

designer knows this then there is no question of providing sound evidence for its trustworthiness to the user.

Evidence that it is worth relying on a given technology depends for its specific content on the aims, interests, and risk-aversity of the user. The specific threshold of functioning and safety that the user demands for justified reliance is particular to her. But there are general kinds of information that can provide a sound basis for trust, that do not require a nuanced understanding of probabilities, uncertainty or statistical reasoning. I focus on two relevant types of evidence: first, evidence that failure to perform will (often) lead to an effective sanction. This makes it clear that it is in the designer's interest to ensure performance. The effective sanction could consist of a punishment, reputational damage, or loss of future opportunities. The evidence need only draw attention to the institutional structures that make such sanctions possible, not to the possibility of specific performance-failures. The second type of evidence consists of an indication that other parties, independent of the designer, and sharing the values of the user, are willing to stake their reputations on the technology's performance. These two types of evidence are sound reasons for trust because they indicate why the designer has a strong interest in serving the interests of the user. They relate both to the likelihood that the technology will perform a certain way, as well as the basis for the user's perceived entitlement to that performance. Russell Hardin describes this as "encapsulation of interests" and thinks of it as the paradigmatic reason for taking the attitude of trust (Hardin 2006). Some social scientific accounts of the reasons of trust, emphasizing reputational and institutional factors, support this view (Coleman 1990; Pettit 1995).

It is highly plausible that in some cases sound information of these types could be made available to the user, and that the user could recognize it as such. Neither form of information depends on difficult, psychologically unreliable forms of probabilistic reasoning.

For example, the user can be shown that an artifact is certified by a governmental authority with the power to sanction failures of performance. At the same time, the second component of trust in technology, the normative expectation, can provide a focal point for communication about what (level of) performance the user should expect. Instead of leaving this implicit, designers can make it clear to the user that certain obvious aspects of performance such as safety and privacy meet established public norms, and it can be made clear to the user what the technology is (and is not) designed to do.

However, so long as a given technology is in fact trustworthy, one might wonder in what way it matters how the designer convinces potential users of this fact. Or to put the point in terms of the user's point of view: so long as she has a true belief that the technology is trustworthy, why is it important that she has access to sound reasons for this belief? One crucial reason is that offered by Plato himself in response to the Meno problem: beliefs held without reason or held on the basis of unsound reasons are brittle, insufficiently "tied down" to ensure a stable attitude. Suppose I decide to join a social networking Internet site, and my reason for trusting the site to keep details about my telephone number, relationship status and political affiliation private is merely that my friends use it regularly and seem to have no such problems. Suppose later one of my friends reports that her private details were made visible to the general public, in just the way I hope to avoid happening in my own case. Suppose that, unknown to me, this happened because my friend made a culpable mistake, but she tells me it is because of a design flaw in the software. If my only reason for believing that the site will keep my details private is the experience of my friends, then at this point my trust may vanish — but if my trust were (also) based on an awareness that an independent body verifies that the site meets established privacy standards, then I might think twice before abandoning my trust in the site on the basis of my friend's experience. Although this is perhaps a trivial

case, we can easily imagine similar cases in which the technology in question was a radiation treatment for cancer, a boat engine, or a new business data management system.

A second reason for providing sound evidence of trustworthiness is that doing so upholds a principle of respect toward users of technology. A widely held ethical principle holds that when important decisions are to be made, those affected by or making these decisions should have the opportunity to give their free, considered consent to the decision, using relevant information. This is known as the Principle of Respect or the Principle of Informed Consent.[12] Although technology users may be limited in their ability to calculate risks decision-theoretically, they have what Gigerenzer and others have defended as "bounded rationality" (Gigerenzer and Selten 2002), allowing them to make rational decisions in many typical contexts. This suggests that, instead of bypassing or ignoring the rational capacities of technology users, it is better to target information to their abilities. Furthermore, some users of advanced technological systems or stakeholders in their use have a better than average ability to make rational judgments about technical reliability and have more knowledge than average about reasonable norms of performance that apply to a given technology. They deserve to have a reasonable opportunity to form a well-grounded attitude of trust or distrust supplemented by evidence about the interests of the technology's designers.

In this chapter I have argued for a distinctive philosophical conception of trust in artifacts and technological systems. I have also argued that this conception invites a new understanding of our way of relying on the designed environment under conditions of risk, and that it allows us to make a distinctive ethical point about technology design. Much more

[12] The principle of informed consent is best known from the domains of research ethics and medical ethics (National Commission 1979). My suggestion here differs from the usual application of the principle of informed consent in three ways: first, I am not suggesting that it be thought of primarily as an institutional or legal norm; second, I am applying it to any technological artifact or system that has significant possible impacts for the user, where the user is in a position to choose that technology; and third, I am applying it to the reasons for trust, rather than to the risks and benefits associated with the technology itself. (In the usual application of the principle, the technology in question would be a research protocol or a medical therapy.)

work needs to be done exploring this topic, including some work that is empirical or interdisciplinary in nature. For example, it would be good to have a better grasp of the psychological basis on which people develop normative expectations of artifact performance, and on how communication affects these expectations. The framework presented here is an *a priori* starting point for such investigations.

References

Aristotle, *The Nicomachean Ethics* (Oxford World's Classics, 1998).

Randolph C. Barrows, Jr and Paul D. Clayton, "Privacy, Confidentiality, and Electronic Medical Records," *Journal of the American Medical Informatics Association* 3 (March/April 1996), 139–148.

Cheryl L.L. Carling *et al.*, "The Effect of Alternative Summary Statistics for Communicating Risk Reduction on Decisions about Taking Statins: A Randomized Trial," *PLoS Medicine* 6, 8 (August 2009), 1–10.

James Coleman, *Foundations of Social Theory* (Cambridge, MA: Harvard University Press,1990).

Paul Faulkner, "On Telling and Trusting," *Mind* 116 (2007), 875–902.

Joel Feinberg, *The Moral Limits of the Criminal Law, vol. 1: Harm to Others* (Oxford UP, 1984).

Maarten Franssen, "Artefacts and Normativity," in Anthonie Meijers, ed., *Handbook of the Philosophy of Science, vol. 9: Philosophy of Technology and Engineering Sciences* (Elsevier, 2009), 923–952.

Lynn J. Frewer, Joachim Scholderer and Lone Bredahl, "Communicating about the Risks and Benefits of Genetically Modified Foods: The Mediating Role of Trust," *Risk Analysis* 23 (2003), 1117–1133.

Gerd Gigerenzer and Reinhard Selten, *Bounded Rationality: The Adaptive Toolbox* (Cambridge, MA: MIT Press, 2002).

Sven Ove Hansson, "Philosophical Perspectives on Risk," Keynote address at the conference *Research in Ethics and Engineering*, Delft University of Technology (2002). Available at http://www.infra.kth.se/phil/riskpage/index.htm.

Russell Hardin, *Trust* (Malden, MA: Polity, 2006).

Pamela Hieronymi, “The Reasons of Trust,” *Australasian Journal of Philosophy* 86 (2008), 213–236.

Richard Holton, “Deciding to Trust, Coming to Believe,” *Australasian Journal of Philosophy* 72: 1 (1994), 63–76.

Karen Jones, “Trust as an Affective Attitude,” *Ethics* 107 (October 1996), 4–25.

Elizabeth A. Lemerise and Kenneth A. Dodge, “The Development of Anger and Hostile Interactions,” *Handbook of Emotions*, 2nd Ed., Michael Lewis and Jeannette M. Haviland-Jones, eds., (New York: The Guilford Press, 2000), 594-606.

David Lewis, *Convention* (Cambridge, MA: Harvard, 1969).

Guido Möllering, *Trust: Reason, Routine, Reflexivity* (Amsterdam: Elsevier, 2006).

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report*. Available at the United States National Institutes of Health Website, <http://ohsr.od.nih.gov/guidelines/belmont.html>. Original report published 1979.

Philip J. Nickel, “Ethics in e-Trust and e-Trustworthiness: The Case of Direct Computer-Patient Interfaces,” Ethics and Information Technology (forthcoming, 2011).

—, “Trust and Obligation-Ascription,” *Ethical Theory and Moral Practice* 10 (2007), 309–319.

—, “Trust, Staking, and Expectations,” *Journal for the Theory of Social Behaviour* 39: 3 (September 2009), 345–362.

Philip J. Nickel, Maarten Franssen and Peter Kroes, “Can We Make Sense of the Notion of Trustworthy Technology?” *Knowledge, Technology and Policy* 23 (2010), 429–444.

Patrick D.T. O’Connor, David Newton, and Richard Bromley, *Practical Reliability Engineering* (John Wiley and Sons, 2002).

Gloria Origgi, *Qu’est-ce que la confiance?* (Paris: VRIN, 2008).

Philip Pettit, “The Cunning of Trust,” *Philosophy and Public Affairs* 24, 3 (Summer 1995), 202–225.

Wolter Pieters, “Acceptance of Voting Technology: Between Confidence and Trust,” in K. Stølen et al., eds., *iTrust 2006, LNCS 3986* (2006), 283–297.

Plato, *Meno*, Trans. W.K.C. Guthrie, in Edith Hamilton and Huntington Cairns, *The Collected Dialogues of Plato* (Princeton, N.J.: Princeton University Press, 1961), 353–384.

Duncan Pritchard, “The Value of Knowledge,” *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), ed. Edward N. Zalta, URL = http://plato.stanford.edu/archives/fall2008/entries/knowledge-value/

J.M. Ramírez and J.M. Andreu, “Aggression, and some related psychological constructs (anger, hostility, and impulsivity): Some comments from a research project,” *Neuroscience and Biobehavioral Reviews* 30 (2006), 276–291.

Ortwin Renn, *Risk Governance: Coping with Uncertainty in a Complex World* (London: Earthscan, 2008).

Fadi Salem, “Enhancing Trust in E-Voting Through Knowledge Management: The Case of the UAE,” in H. Qian, M. Mimicopoulos, H. Yum, eds., *Managing Knowledge to Build Trust in Government*, (New York, United Nations Department of Economic and Social Affairs (UNDESA), 2007).

Mariarosaria Taddeo, “Trust in Technology: A Distinctive and a Problematic Relation,” *Knowledge, Technology and Policy* 23 (2010), 283–286.

Amos Tversky and Craig R. Fox, “Weighing Risk and Uncertainty,” *Psychological Review* 102 (1995), 269–283.

Amos Tversky and Daniel Kahneman, “Advances in Prospect Theory: Cumulative Representations of Uncertainty,” *Journal of Risk and Uncertainty* 5 (1992), 297–323.

Anton Vedder, “Betrouwbaarheid van internetinformatie,” in J. de Haan en J. Steyaert, eds., *Jaarboek ICT en samenleving 2003: De sociale dimensie van technologie* (Amsterdam: Boom/SCP, 2003), 113–132.

Pieter E. Vermaas, Yao-Hua Tan, Jeroen van den Hoven, Brigitte Burgemeestre, and Joris Hulstijn, “Designing for Trust: A Case of Value-Sensitive Design,” *Knowledge, Technology and Policy* 23 (2010), 491–505.

Mattias J. Viklund, “Trust and Risk Perception in Western Europe: A Cross-National Study,” *Risk Analysis* 23 (2003), 727–738.

Leif Wenar, “Rights,” *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/rights/>.

Ludwig Wittgenstein, *On Certainty* (New York: Harper, 1969).

Crispin Wright, “On Epistemic Entitlement,” *Proceedings of the Aristotelian Society Supplementary Volume* 78 (2004), 167–212.