Know Techn Pol (2010) 23:429–444
DOI 10.1007/s12130-010-9124-6

SPECIAL ISSUE

Can We Make Sense of the Notion of Trustworthy Technology?

Philip J. Nickel · Maarten Franssen · Peter Kroes

Received: 1 June 2010 / Accepted: 26 July 2010 / Published online: 17 September 2010

Abstract In this paper we raise the question whether technological artifacts can properly speaking be trusted or said to be trustworthy. First, we set out some prevalent accounts of trust and trustworthiness and explain how they compare with the engineer's notion of reliability. We distinguish between pure rational-choice accounts of trust, which do not differ in principle from mere judgments of reliability, and what we call "motivation-attributing" accounts of trust, which attribute specific motivations to trustworthy entities. Then we consider some examples of technological entities that are, at first glance, best suited to serve as the objects of trust: intelligent systems that interact with users, and complex socio-technical systems. We conclude that the motivation-attributing concept of trustworthiness cannot be straightforwardly applied to these entities. Any applicable notion of trustworthy technology would have to depart significantly from the full-blown notion of trustworthiness associated with interpersonal trust.

Keywords Artificial intelligence · Trust · Trustworthiness · Technology · Socio-technical systems

1 Introduction

Philosophers may have difficulty in spelling out exactly what it means to have trust in persons or persons being trustworthy, but *that* it makes sense to apply the notions

P. J. Nickel (✉)
Eindhoven University of Technology, Eindhoven, Netherlands
e-mail: p.j.nickel@tue.nl

M. Franssen
Delft University of Technology, Delft, Netherlands
e-mail: P.M.Franssen@tudelft.nl

P. Kroes
Faculty Technology, Policy and Management, Delft University of Technology, Delft, Netherlands
e-mail: P.A.Kroes@tudelft.nl

of trust and trustworthiness to people is not controversial. Any theory of trust and trustworthiness that deserves to be taken seriously should be able to explicate what it means to have trust in persons or for persons to be trustworthy. Furthermore, such a theory must be a fruitful basis for explaining the role of trust in interpersonal behavior, such as situations of (social) cooperation. However, the situation is different when we try to apply notions of trust and trustworthiness to technical artifacts or technical systems. Does it make sense at all to say that we trust the navigation system in our car or that it is trustworthy? Or that we trust a shopping site on the internet, or that modern technology in general is trustworthy? Is this just a manner of speech, a metaphor that will not survive closer scrutiny because in saying that a technical artifact is trustworthy a category mistake is made? If not, if a genuine sense of trust and trustworthiness is involved, then the important question arises whether there are significant differences between trust in (trustworthiness of) persons and trust in (trustworthiness of) technical artifacts. In that case, moreover, there is the question how the trustworthiness of technical artifacts is related to their technical reliability.

The increasing tendency to speak about technology as trustworthy and about people trusting technology fits well within various developments in which the traditional conceptualization of technology as a collection of passive, inert instruments is called into question. Starting from the idea that technology affects human beings in various kinds of ways, that technology *does* all kinds of things to human beings in apparently more or less autonomous ways (see, for instance, Verbeek (2005)), there has been an increasing tendency in Science and Technology Studies and in the ethics of technology to conceive of technology as having some form of agency of its own. In line with this technical artifacts are claimed to have politics (Winner 1985; Winner 1992 (1977)) or a moral status or agency of their own (Latour 1992; Latour 2002). Independent of these developments, the ever increasing use of artificial intelligence in technical artifacts has resulted in a situation in which in many cases the behavior of technical artifacts comes close to human behavior. As a result, such technical artifacts are often characterized as 'artificial' agents. What these various developments have in common is that notions that play a key role in describing and explaining intentional human behavior, such as "autonomy", "making choices", "agency", and "morality", are being used to describe and explain the behavior of technical artifacts as products of intentional human behavior. Ascribing trustworthiness to technical artifacts is another example of this tendency to attribute features to technical artifacts that were originally applied primarily or exclusively to human beings. One of the arguments used to justify these ascriptions is to claim that technical artifacts, as products of human intentional action, have some form of derived intentionality (Ihde 1990; Latour 1992; Huyke 2003; Verbeek 2008).

In this paper, we will examine whether and, if so, under what conditions a notion of trustworthiness may be attributed meaningfully to technical artifacts. We start by looking at trust in persons: reliability and trustworthiness are distinct from one another even as applied to persons. In Section 2 we describe two contrasting accounts of interpersonal trust—pure rational-choice accounts and motivation-attributing accounts—as a way of exploring the characteristics of trustworthiness as an attribute of trusted people. In Sections 3 and 4 we extend these accounts of trust and trustworthiness to technical artifacts, starting with a discussion of how engineers define the reliability of technical artifacts and human reliability. In Section 5 we discuss two examples of technical artifacts for which it seems to make sense to claim that they are the objects of trust and may be trustworthy. In Section 6 we turn to a discussion of the notion of trust in socio-technical systems, such as electric power supply systems, public transport systems, and e-commerce systems. Here we are dealing with the notion of trust in systems composed (at least) of technical artifacts and human beings. Although it is common practice to speak about trust in, for instance, governmental systems the use of this notion of trust has been criticized (Hardin 2006, p. 41). Our analysis focuses on the role of trustworthiness of operators in socio-technical systems. In the final section we summarize our main conclusions.

2 Accounts of Trust in and Trustworthiness of Persons

The concepts of trust and trustworthiness are complex and have been subjected to much scholarly interest across many academic disciplines. Here we will confine our attention to trust considered as a three-place relation between a trusting agent A, some trusted entity B, and an anticipated, desirable performance *x* (a formulation going back to Horsburgh (1961)). In the case of interpersonal trust, B is a person and *x* is an action performed by B. Later in the paper, we will address a different case, in which B is an artifact or a socio-technical system and *x* is an anticipated event caused by B. Our reason for focusing on three-place trust is that it is central to the explanation of particular actions performed on the basis of trust. A's simple "two-place" feeling of trust in B, insofar as it does not concern particular performances of B, does not play the same fine-grained role in explaining why A relies on B in specific ways or contexts. We will also consider trustworthiness to be a three-place relation: B is trustworthy for person A with regard to the performance of *x*. Again, in interpersonal trust B is a person, but when talking about trustworthy technology, B would be a technical artifact or a socio-technical system.

The simplest account of trust is a pure rational-choice account, conceiving of trust as a judgment one makes when, having considered the possible costs and benefits of relying on another person's performance, as well as the salient alternatives, one comes to the conclusion that it is *worth it* to rely on that performance (Coleman 1990). Such an account fits squarely with a conception of action as a (subjectively) rational choice, a choice made on the basis of one's perception or estimation of the expected consequences. It must be stipulated that the reliance in question is free from coercion, and that the performance relied upon is desirable in some way, or else the account would be forced to regard willingness to rely on others in circumstances of threat or calamity, as well as reliance on B's unwanted actions, as cases of trust. Indeed it is only by stipulation that the pure rational-choice account distinguishes between interpersonal trust and reliance on natural events. The need for these stipulations shows that the pure rational-choice account does not make any principled distinction between trust and mere rational reliance. The account, therefore, is unfit to explain why performance-failure in cases of genuine trust leads to appropriate feelings of blame or betrayal in cases of malevolent breach of trust, and directed anger or disappointment in cases of negligence or incompetence (Nickel 2009).

For this latter reason, many accounts of trust regard attributing certain *motivations* to the trusted person as essential for genuine interpersonal trust. These attributed motivations come in many flavors: one expects the trusted person to possess moral integrity (McLeod 2002); one holds the trusted person to a moral obligation (Nickel 2007); one believes that she is aware of and cares about normatively salient features of the situation such as the interests of the truster (Hardin 2006); or one counts on her to recognize the very fact that the truster depends on her as a reason to perform (Jones 1996; Faulkner 2007). Various though they are these motivation-attributions provide a principled distinction between genuine trust and a simple judgment that reliance on the behavior of another person (or a natural event, for that matter) is worthwhile or rational. They will be important later for considering trust in technology.

Trust and trustworthiness are interlocking but categorically distinct (Hardin 2006). Trust is usually taken to be an attitude, whereas trustworthiness is a quality in the object of this attitude that satisfies it and helps make it appropriate.[1] In trusting B to do *x*, A is always committed in principle to regarding B as trustworthy enough to warrant trust. But in practice, pragmatic factors such as the availability of other options besides relying on B, and the desired extrinsic effects of reliance on B (e.g., effects on the relationship between A and B, or on B's development of new competencies), are also crucial in coming to trust (Holton 1994). Trustworthiness, conversely, is that quality which provides sufficient reason to warrant the attitude of trust. In other words, if someone is trustworthy for person A with regard to *x* then it is rational for A to trust that person with regard to *x*. For the pure rational-choice account, this consists straightforwardly of exhibiting predictable behavior of the desired type. And certainly people can have such a property of reliability. But since the rational-choice account fails to capture a crucial feature of the attitude of trust, the mere quality of predictably acting in a favorable way is not sufficient for genuine trustworthiness. For example, Immanuel Kant is said to have gone on walks at precisely the same time each day. Others who relied on this act of Kant's to set their watches surely cannot be said to have trusted him in the relevant interpersonal sense (Baier 1986, p. 235).

If, instead, we take motivation-attributing accounts of trust as our guide to an account of trustworthiness, we should analyze trustworthiness partly as consisting of having certain motivations or capacities such as taking the interests of others into account, or taking into account the fact that others are relying on one. According to various motivation-attributing accounts of interpersonal trust, trust will consist respectively of: possessing moral integrity; being capable of fulfilling one's moral obligation; actually caring about the interests of the truster; or, recognizing and responding positively to the very fact that one is being relied upon. Since all of these qualities are normally possessed only by humans, it would make sense to conclude that technology can never be genuinely trustworthy; it can only be reliable. More in particular, from the perspective of motivation-attributing accounts of trust,

[1] Note that whereas trust (trusting somebody) is usually taken to be an attitude, putting one's trust in somebody is not an attitude but a (deliberate) action. In the following we will confine our analysis to trust and sidestep any issues concerning differences between the notions of trust(ing) and of putting trust in. (It is conceivable that someone might put her trust in a person without trusting that person).

trustworthiness implies that for something to be trustworthy that something is capable of having mental states that have as their content the values and interests of other entities and that the trusted thing has interests of its own. Since only individual persons qualify as having mental states and as having interests, this implies that both the trustworthy entity and the trusting entity are individual persons. Consequently, a direct application of motivation-attributing accounts of trustworthiness to technical artifacts (technology) seems out of the question. The transference of the interpersonal notion of trustworthiness to technical artifacts will require some modifications of the motivation-attributing account. In the following we will explore what kind of obstacles hinder the application of motivation-attributing accounts to technical artifacts. To that end we first turn to a clarification of the notion of reliability of technical artifacts and persons from the perspective of engineering practice.

3 Extension of Accounts of Trustworthiness to Technical Artifacts

The following definition of reliability epitomizes how engineers think about reliability (Birolini 2007, p. 2):

> "*Reliability* is a *characteristic* of an item, expressed by the *probability* that the item will perform its *required function* under *given conditions* for a *stated time interval*. It is generally designated by *R*. From a qualitative point of view, reliability can be defined as the *ability of the item to remain functional*. Quantitatively, reliability specifies the *probability that no operational interruptions* will occur during a stated time interval."

Note that according to this definition reliability is not an evaluative notion, neither quantitatively or qualitatively. It does not refer to any norm or standard for reliability. The reliability of a technical artifact is simply one of its empirical characteristics or features that can be measured in the defined way. Strictly speaking, on this definition it does not make sense to say that a technical artifact is reliable. Such an absolute statement makes an implicit appeal to a normative threshold for reliability, such that when the artifact meets this threshold it is called reliable. Usually reliability is taken to be a desirable feature of technical artifacts: all else being equal, the more reliable technical artifact is to be preferred to the less reliable one. However, this does not mean that one should strive unconditionally for maximum reliability. If all else is not equal, trade-offs may have to be made between reliability and, for instance, efficiency, durability, and cost, which implies that in particular cases a less reliable technical artifact might turn out to be the preferred one.

Although the above definition is primarily intended to be applied to technical artifacts, we may take the item to which it refers to be a person, in which case we end up with the following definition of reliability of a person:

> *Reliability* is a *characteristic* of a person, expressed by the *probability* that the person will perform his/her *required function* under *given conditions* for a *stated time interval*. ... From a qualitative point of view, reliability can be

> defined as the *ability of the person to remain functional*. Quantitatively, reliability specifies the *probability that no operational interruptions* will occur during a stated time interval.

From an engineering point of view this definition of reliability of a person makes perfect sense when that person is the operator of a machine. Just as the machine has a particular function, so the person *as the operator* of the machine has a function or role and the reliability of that person/operator may be defined in terms of how well it performs that function or role. Furthermore, just as a machine may malfunction, so an operator may malfunction (may make errors in the performance of an operator's role).

The above definition of reliability of a person is roughly in line with the way Whittingham (2004) explicitly defines human reliability in the context of analyzing human errors as causes of accidents. According to Whittingham (2004. p. 46) the most common way "to quantify human reliability in a given situation is that of human error probability (HEP)" where HEP is defined as the mathematical ratio between the number of human errors occurring in a task and the number of opportunities for human errors. Human error is defined in the following way (Whittingham 2004, p. 6):

> "A human error is an unintended failure of a purposeful action, either singly or as part of a planned sequence of actions, to achieve an intended outcome within set limits of tolerability pertaining to either the action or the outcome."

Human reliability now equals one minus the HEP. By assuming that a human error corresponds to an operator malfunctioning, the parallel between the two definitions becomes immediately clear.

The foregoing illustrates that from an engineering point of view there appears to be no fundamental difference between the reliability of machines and people, in so far as the latter are taken to be operators (we will come back to this point in Section 6). The importance of the notion of reliability for engineering practice, and for users, is rather obvious. Knowledge of the reliability of a technical artifact or an operator forms the basis for predictions about the expected success of the use of a technical artifact or operator.

Let us now turn to an examination of how the various accounts of trust may be extended to technical artifacts. The rational-choice account of trust can be extended to technical artifacts in such a way that we end up roughly with the engineer's conception of reliability of technical artifacts. This works as follows:

- Since we are interested in trust in technical artifacts we allow that the trusted thing (the thing that we rely upon for the desired performance) is a technical artifact rather than a person.
- We assess the likelihood of the desired performance of a technical artifact using the standard probabilistic measurement scale ranging from zero to one.
- We assume that trust in a technical artifact can be measured on a quantitative scale; in other words, we assume that trust is not an all or nothing affair. Trust comes in degrees ranging from complete distrust (corresponding to the value zero) to complete trust (value 1).

Adapting the rational-choice account of trust accordingly, we end up with the following account of trust in technical artifacts: "one trusts a technical artifact to a degree *x* when and only when one is willing to risk the use of that technical artifact on the basis of a judgment that it will perform (function) with probability *x*." Now, since performance (functioning) with a probability *x* is the core idea underlying the engineer's notion of reliability, what we end up with here is an account of what it means to trust (to a certain degree) a technical artifact that is more or less identical with what it means to rely (to a certain degree) on it.

So, we reach the conclusion that extension of the rational-choice account of trust to technical artifacts does not lead to a genuine notion of trust in technical artifacts, different from reliability. Taking into consideration that engineers treat technical artifacts and persons, in so far these are operators, with regard to reliability in the same way, our conclusion may be taken as an indication that in the rational-choice account of trust in persons, persons figure more or less as operators who are expected to perform a task.

Turning our attention to motivation-attributing accounts of trust, we observe that they cannot be applied straightforwardly to artifacts. For example, Hardin's encapsulated interest account of trust (Hardin 2006), which straddles the line between a rational-choice account and a more psychologically complex, motivation-attributing account of trust, is difficult to apply to artifacts. Hardin also takes trust to be a three-place relation: A trusts B with regard to *x*, where *x* stands for certain actions that are predicted to a certain degree from B, and where B's role is to perform those actions.[2] This means that B is treated more or less as an operator in Hardin's account, and we may end up with the engineer's idea of the reliability of operators. However, the idea of the encapsulated interest account of trust also imports something new, over and above this rational assessment of likelihoods: namely, certain motivations or reasons that are expected to play a part in causing B to do *x*. The relevant reasons must include the idea that the trusted person has one's own interests among his interests. This is why Hardin says that the encapsulated interest account "is a rational expectations account in which the expectations depend on the *reasons* for believing that the trusted person will fulfill the trust (p. 31)." Since technological artifacts and systems (apart from not being persons) cannot be said to have interests in any straightforward sense, much less to have others' interests encapsulated in their own, Hardin's account seems unpromising as an account capable of grounding the notion of trust in technology.

Moralized notions of interpersonal trust, such as that in Nickel (2009), according to which trust implies a moral expectation that the object of trust will perform a certain way, also cannot be adapted straightforwardly to technology. In order to ascribe a moral obligation to some entity, one has to suppose that the entity is capable of responding to moral considerations and performing with responsibility and discretion. *Simple* technological artifacts cannot be said to take moral considerations into account, or to be responsible, or to have discretion in how they perform. Before we turn to an examination of trustworthiness of *complex, agent-like*

[2] Misunderstandings about the role and the discretionary powers that come with the role may lead to betrayal of trust; see Baier's (1986) discussion about the babysitter.

technical artifacts, we analyze trust in technical artifacts as a derived form of interpersonal trust.

4 Trust in Technical Artifacts as a Derived Form of Interpersonal Trust

If it is understood that in order to trust something, or for something to be trustworthy, this something must not only have mental states[3] but must have mental states in which the interests of the trusting person are represented, then the notion of trust in technological devices cannot make sense (supposing that no technological device is currently a candidate for having mental states ascribed to it, a situation that may change in the future). However, technology is not just a collection of devices. People are involved in technology essentially, and in various ways. First of all, technical devices are designed, manufactured, sold, maintained and, if necessary, repaired by people. This creates a difference between using a bridge to cross a river and using a tree that happened to fall across the river to do so. In this section we discuss whether this allows us to speak of trust in (forms of) technology, if not straightforwardly then perhaps in a derived sense, and whether this move can ultimately be recommended.

Suppose that someone—let's call him Floris—buys a new kind of coffee machine or some other consumer product and plugs it into a socket. Typically, Floris is confident that it will not blow up in his face. In deciding to use it, he relies on the device doing what the documentation says it will do and not something else that will harm him. The status of something as a technical artifact is particularly associated with the reason for one's reliance on it. Suppose Floris uses some object, be it natural or artificial, for an incidental purpose—say something heavy for hammering a pole into the ground or for temporarily supporting some other thing. In this case his reliance that things will go the way he intends them to go, assuming his reliance satisfies minimal requirements of rationality, is based on his *confirmed* judgment that the object has the material properties that enable it to serve his purpose in the way he intends. The crucial word is 'confirmed': Floris can use the object because he has checked for himself that the thing has the required properties, possibly by indirectly inferring the presence of these properties from other properties he knows the thing to have, or possibly by seeing someone else use the object in the same way. In contrast, when Floris buys a new kind of coffee machine and plugs it in, he has often no idea what material properties it must have in order to do what it is supposed to do, let alone that he has checked that the product has these properties. Floris's confidence concerning the product's expected operation must therefore be grounded in some attitude of his concerning the people who made the device or sold it to him. It is difficult to see what other ground his confidence could have.

But is this confidence an attitude of trust? If his newly bought coffee machine does blow up when plugged in, certainly he is not merely disappointed (supposing

[3] Some accounts of trust, in conceptualizing relations between artificial (electronic) agents, do not see trusting itself as being based on a trustor's mental state. See, for example, the accounts of e-trust in Taddeo (2010) and De Paoli and Kerr (2008). Since our discussion is targeted toward trustworthiness rather than trust, and is not focused on electronic trust networks among artificial agents, in this paper we do not undertake to relate non-mental notions of e-trust to more traditional theoretical conceptions of trust.

he survives)! He will feel this should not have happened, even if it was not anybody’s intention it should. Floris cannot, of course, expect the designer/manufacturer to have mental states concerning *his* interests or feelings in particular. But they can be expected to know that whoever buys one of their coffee machines, that person is not likely to appreciate it exploding. If it is sufficient for someone to be trustworthy that this person has an interest in building a reputation and will therefore take one’s interests into account, then Floris may trust the designer/manufacturer of his coffee machine to put a product on the market that has an extremely low risk of spontaneous explosion. Suppose that this exhausted Floris’s reasons. Then, it would depend on the sheer number of products they sell, on the costs of avoiding the misbehavior of a product, on the costs to their reputation when a product does misbehave, and on the likelihood of such an occurrence actually damaging their reputation (which would depend again on the quality of the media, of the legal system, and so forth) how much effort would be spent on taking care that Floris's machine should not blow up when plugged in. And when it blows up all the same, Floris may be reminded of the fact that the producing company has made such calculations. He is likely to be disappointed or frustrated. But would he go so far as to say that his trust has been betrayed? Not all disappointments about behavior are betrayals of trust. For a betrayal of trust, something more must be the case than a mere expectation that someone will do X because that person is taking your interests into consideration since it is *rational* for that person to take your interests into consideration. The something more is typically that this person has let you know, promised you even (at least implicitly), that he or she will act in accordance with your interests.[4] Betrayal is, in typical cases, nothing else but the breaking of an (implicit) promise.

Now in technology, the situation is arguably like this. Advertising a product as a coffee machine, adding to it an instruction manual showing you how its manipulation will result in a cup or a carafe of coffee, amounts to promising you that the product you bought will make coffee when plugged in (and not in a perverse way, i.e., by blowing up in your face while making coffee at the same time). An interesting case is one where conflicts of interests arise: where a promise is made that it is not, invariably, rational for the promise-maker to keep. It may be that corporate promises fall, *ipso facto*, in this category.[5] The fact that companies have introduced such things as guarantees may show that they realize that the reliability of the promises they make concerning their products is questionable. However, the contrast with the promises of individuals is easily exaggerated. Individuals are not less regularly tempted to reconsider their promises in the light of their own interests, and a promise without a reputation to support it does not easily lead to trust. The reason that a betrayal of trust is experienced as a significant event may well be that it represents a double failure: a failure on the part of the trustee to live up to a promissory expectation and a failure on the part of the truster to anticipate the

[4] There are exceptions. It is possible that mutual but uncommunicated assumptions make it reasonable to assume the other person will take your interests into account *qua* your interests, even though the person has not specifically communicated this.

[5] Cf. Hardin’s skepticism concerning trusting the government, mentioned in Section 1.

contrary outcome. As the very word indignation indicates, the indignation felt when trust is betrayed is a combination of anger and shame.

Technical artifacts are in fact embedded in a network of promises: they are designed, manufactured, and sold as specimens of functional kinds; as things that are for being used in certain ways and thereby realizing certain outcomes, and that may therefore be expected to perform accordingly. These promises themselves, however, cannot ground trust concerning the artifacts' operation. Many promises are made about second-hand cars, and they do not resolve the question whether the seller can be trusted; they make it more poignant. The promises make it the case that the disappointment of our trust concerning an artifact's performance is also a betrayal of trust. In conditions where the producers seem not to be hindered by a bad reputation, and do not care about their reputation, as was typically the case under the former communist regimes, people are much less inclined to trust promises about artifact's performance. If people are disappointed under such circumstances, it is disappointment about the failure of a plan, the sort of disappointment felt when a lottery ticket turns out not to have won a prize.

These considerations concern trusting the designers and manufacturers of technical artifacts, not the artifacts themselves. We could, however, introduce trust in artifacts in a derived sense. To trust an artifact to perform in a certain way would mean, then, to trust the designer or manufacturer of the artifact to have seen to it that the artifact performs, as we might put it, in accordance with the description of its function and only in accordance with the description of its function and operation, as communicated by the designer/manufacturer. If we trusted an artifact in this derived sense, we would consider it reliable.[6]

From the perspective of this paper the claim that technical artifacts are trustworthy in this derived sense is not very interesting since it traces the trustworthiness of technical artifacts back to the interpersonal trustworthiness of designers, makers etc. What we are interested in here is whether the notion of trustworthiness may be applied to technical artifacts themselves, that is, whether a non-interpersonal notion of trustworthiness of technical artifacts may make sense. We have already concluded above that it is not at all clear how motivation-attributing accounts of trustworthiness may be adapted such that they may be applied to simple technical artifacts. In the following section we explore to what extent it is possible to apply motivation-attributing accounts to more complex technological objects of human reliance, in particular to technical objects that show agent-like behavior. We discuss the question whether they can show sufficient agent-like behavior to count as being in some sense trustworthy, and as suitable objects for trust. In such cases, however, we are dealing with 'thin' counterparts to the notions of trustworthiness and trust found within motivation-attributing accounts.

[6] The reverse would not be true. So long as we exclude from trust the motivation-attributing thought that among the reasons of the designer/manufacturer for delivering an artifact that performs as advertised are considerations about our interests in the artifact performing in that way, we can consider an artifact reliable while untrustworthy. But is there any point in excluding this idea from trustworthiness? It is difficult to see any work that such a notion of trustworthiness could do, in addition to the work that the notion of reliability already does.

5 Examples: Trustworthy Navigation Systems and DCPI's

Let us start by considering two examples of technological artifacts with features that make them plausible objects of trust-attitudes: on-board vehicle navigation systems, and direct computer–patient interfaces (DCPIs). Both kinds of technological artifacts display a sort of embedded intelligence, but it does not make literal sense to say that they have intentions, beliefs, wills, or interests of their own. The key features that make them suitable for trust are that they:

1. Use natural language to interact with human users;
2. Operate in contexts where risk and other morally significant features are present;
3. Adapt their activity to the context; and,
4. Incorporate values into their performance.

Here are two vignettes of user interaction with these artifacts that illustrate these trust-suitable features:

> NAV
> Doris, a non-expert user of navigation systems, is driving to a job interview 60 km away from her home. Coincidentally, her sister also commutes daily along this road. Normally, the road is not busy, but today there is bad weather and construction, and Doris risks running late. Fortunately, she thinks, she has an on-board navigation system, NAV, that can help. NAV is designed to promote sustainable driving behavior by selecting the route most likely to conserve fuel; it is continuously updated with the latest news about traffic jams, construction works, etc. Although the system can in principle select the fastest route, this is not prioritised in the design. In light of the circumstances and in order to save fuel, NAV advises Doris to take a route off the main motorway (using speech synthesis), and she follows this advice. Doris arrives late to her interview. Later, Doris learns from her sister that the weather and construction delays were not as bad as anticipated that morning, and that she likely would have been on time had she stayed on the main motorway.
>
> DCPI
> Boris lives in a rural area far away from medical specialists and has limited mobility because of a physical disability. Recently, his only living relative, his mother, passed away, leaving him a small amount of money which he used to purchase a computer. One of his friends, worrying that he has been depressed since his mother's death, advises him to complete an online diagnostic questionnaire about depression. The online questionnaire was developed by psychiatrists to reach rural patients who might otherwise not seek medical help for psychological illness. After answering a sequence of questions about his mood and symptoms, the online system diagnoses Boris with likely depression. It also leads Boris through a tutorial on types of antidepressant medications that could be used for his condition and their effects. Finally, the online system advises Boris to take some antidepressant medication that he orders from a company through the Internet. The medication helps with his mood but causes sleep loss and high blood pressure.

Intuitively, it seems meaningful to ask whether Doris and Boris trust the respective technological artifacts they are using in these two vignettes. This may involve more than just the question of the systems' reliability, depending on how well the system adjusts itself to the values, including ethical values, that Doris and Boris consider relevant to their respective situations. In the case of NAV, Doris may consider it a priority to be on time to appointments even while sacrificing fuel conservation. Although she recognizes energy conservation as having moral weight, she may think there are important exceptions to its overridingness. NAV caused her to be late because its settings were based on considerations of saving fuel. This makes it unreliable for Doris in a situation where she values time more highly. But we can imagine a variant of the case in which the fuel-saving mode is merely the system's default setting, and that it has an alternative urgency mode, to which it switches when it observes physiological states of stress and impatience in the driver, or that it asks the driver to rate the importance of arriving on time. If such a setting had been operating in the NAV case, Doris's reliance would have been rewarded. She might have been impressed at how the system adapted itself to her value preferences, taking her interests into account. In such cases, artifacts appear to the user to encapsulate her interests and share her values. Likewise, Boris might have trust-like attitudes toward the DCPI to the extent that it reflects his own interests or ethical judgments. For example, he might be hesitant at first to accept the judgment that his symptoms are signs of an illness requiring treatment, rather than a normal reaction to his mother's death. His attitude toward the technology may ultimately depend on whether he accepts the alternative ethical judgment that the DCPI appears to endorse. Again we can imagine that the system shows itself sensitive to Boris's interests, by asking questions about his attitude to medication, or by telling him, when the final diagnosis is presented, that the symptoms are serious enough to advise him to consult a physician.

But if we speak of trust and distrust here, it may be more appropriate to interpret it as being directed at the designers of the DCPI, rather than the artifact itself. The designers might, for example, partly represent the interests of the pharmaceutical industry, thwarting the interests of users like Boris by overdiagnosing illnesses requiring medication. If so, it is the designers who should be the target of moral disapprobation. Thus it remains an issue whether incorporating ethical values really makes NAV and DCPI candidates for being trustworthy rather than just reliable. Something still seems to be missing. Even when such systems acquire capacities that come close to taking the interests of users into account, they still lack something that seems to be required for the notion of trustworthiness to apply: interests of their own that could trump the interests of the users. This is crucial for the feeling that these systems have let the users down rather than merely that they have shown themselves to be unreliable tools.[7]

[7] See, however, Nickel (2011) for a 'thin' account of trust in artifacts that construes it as a non-moral but nonetheless object-directed expectation.

6 Operators and the Trustworthiness of Socio-technical Systems

If trust requires the taking into account of interests and the possibility of a conflict of interests, then trust in technology must be derived in some way from interpersonal trust. It is important, therefore, to investigate the different ways that people are involved in technology. Technology is not just a collection of devices intermediating between their designers on the one hand and their users on the other. Many systems used by people to achieve their purposes are not purely technical devices; they include people in various roles. Such systems are *hybrid*, partly technical and partly intentional or social, and can therefore be called socio-technical systems. If Doris uses a car to drive to her job interview, then she uses a mechanical device. It is entirely up to her to arrive where she wants to go by manipulating the car's wheel, handles, pedals, and buttons. When she takes the bus or train, however, she is using something that includes both the vehicle and its driver. It is the driver who controls the vehicle; the manipulations required of Doris to arrive at her destination are of a completely different sort. And even when driving herself, she uses roads, lanes, exits, traffic lights, and other items, the precise 'manipulation' of which depends on the presence of other drivers, pedestrians, and the occasional police official, and perhaps on other people whose presence she does not notice but who control the various indicators for maximum speed, lane access, and so forth.[8]

The network of advertised functions serving as promises discussed in Section 4 extends to socio-technical systems, where what is delivered to us is not a physical product that we ourselves then use, but where all physical devices are components of the system we use and what is delivered is a *service*. We would have little reason to trust the supplier of transportation or electric power to live up to its promises of reliability if we did not assume that this supplier cared about its reputation. Still, we trust the suppliers or owners of the system delivering the service, not the system itself, just as we trust the designer/manufacturer of a product, but not the product. Service-providing socio-technical systems, however, have numerable components, and many of these components are people, to whom we shall with a general term refer as operators. Operators have mental states and are capable of considering the interests of the users of the system and how their actions as an operator may affect these interests. By law operators are often obliged to consider these interests, even if it is possible to perform the specific operator role without giving these interests any thought and the operator role can be defined without referring to them at all. By law, for example, the driver of a bus and the pilot of an airplane are responsible for its passengers while performing their operator roles of driving the bus or flying the plane.

The users of a transportation system involving buses will want the bus driver to be trustworthy in the performance of her role as a driver, meaning that she will not let the passengers' interest be trumped by either her personal interests as a human being or her professional interests as a system operator and, typically, company employee. An example of the former would be that the passengers may expect the driver to drive smoothly although driving roughly would be more fun. An example

[8] For a detailed exposition of hybrid systems, which can be decomposed into roles some of which are fulfilled by humans and some by material devices, see Franssen and Jespersen (2009).

of the latter would be that the passengers may expect the driver not to take major risks in order to keep up with the system's time schedules. A major accident with a commuter train in Amagasaki, near Osaka, Japan, on 25 May 2005, caused by the driver's reckless speeding out of fear for lagging behind the company's time schedule, shows that this is a live issue.

System users will therefore also want to be able to trust higher-level operators not to set conditions for lower-level operators that will cause these lower-level operators (such as bus and train drivers) users' interests to be trumped more easily. It is of course an issue what minor risks the users may expect the driver to take, because keeping to the time schedule within reasonable limits is among their interests. A case where a mixture of all these considerations may have been at work is the notorious Tenerife airport accident where an airplane at take-off crashed into an airplane taxiing on the runway. The crew of the first airplane had been overhasty in taking off out of fear of running so much behind their schedule that the whole flight had to be canceled, which would mean a great nuisance to the passengers and to themselves as private persons and additionally cause significant costs to the airline.

Even if we have trust in operators of the system, this does not mean that it makes sense to say we trust the system. That still remains an open issue. The system as such does not take the interests of its users into account. In fact, the more operators there are, the less coordinated their actions and considerations will be with respect to the interests of users. As in the case of the reliability of technical artifacts, the trustworthiness of the system's operators merely diversifies the reasons its users have to consider the system reliable. In this, it is typical for socio-technical systems that the engineers who design them are distrustful of operators being candidates for trustworthiness. The capacity that people have to entertain a broad spectrum of considerations makes it difficult to design such systems for controllability, which is an engineer's main goal. If an operator makes his inclination to act according to the instructions that define his role dependent on additional considerations of interests, either his personal ones or the interests of particular users, the reliability of the system becomes dependent on the ability of the operator to choose the optimal balance between these interests time and again. Automation is the engineer's answer, through building engineered hardware systems that will perform better, in the long run, than a human operator trying to balance an uncontrollable and often idiosyncratic spectrum of interests.

The users of socio-technical systems, for their part, tend to decrease opportunities for engineered control, bypassing regulations, usurping operator prerogatives, and emphasizing individual freedom instead. For example, the condition of the road traffic transport system, in contrast to the air and rail transport systems, is to a large extent determined by the behavior of individual drivers. Many of the reasons we have for trusting the operators in the latter systems are absent in the former. Individual drivers are not much restrained by considerations of reputation: road traffic is more like a sequence of one-shot prisoner's dilemmas than a sequence of repeated prisoner's dilemmas between identical players with an indefinite time horizon. Many individual car drivers seem to care more about their own lives than about the lives of other drivers, judging from the recent 'arms race' towards heavier and heavier SUVs. Apart from safety considerations, the reliability of road transport systems is also determined by congestion issues, and here as well the impulsiveness of individual drivers and the lack of coordination among their actions is a major cause of loss of reliability. On the road, then, we can do little more than put our trust on other drivers as we attempt to get to our respective destinations.

By making systems more reliable, engineers shift the focus of the user's trust away from operators as system components to the designers of the system, and make our trust less evenly distributed, so to speak. In this way they act to decrease opportunities for trusting system operators, replacing them with automated devices. When all operators have been replaced by hardware–software devices, a hugely complex technical artifact results. Only the notion of reliability applies unproblematically to this artifact, whereas the notion of trustworthiness raises problems. Automation deprives the users of a system of the reasons they have for assessing a system's reliability by way of trusting its operators. It may well be that while the system is actually made more reliable, it is perceived as becoming less reliable. Insofar as the trustworthiness of socio-technical systems can be defined, it is likely to be in terms of the trust people have in its operators, but because the presence of these operators may compromise reliability, such a definition of trustworthiness could lead to conflicting judgments of reliability and trustworthiness.

7 Conclusion

On a pure rational-choice account of trust and trustworthiness, the trustworthiness of technical artifacts turns out to be the same as the reliability of technical artifacts. Such an account does not lead to a genuine notion of trustworthiness of technical artifacts distinct from their reliability. On the other hand, motivation-attributing accounts of trust and trustworthiness cannot be adapted in a straightforward way to simple technical artifacts since these accounts presuppose that the trustworthy technical artifact has (1) mental states, (2) the content of which is about the interests and values of the trusting person, and (3) has interests of its own. Complex technical artifacts exhibiting agent-like behavior might be called trustworthy in a sense that is different from reliability. However, we argue that this would have to be a thin notion of trustworthiness. It would differ significantly from the full-blown, motivation-attributing notion of trustworthiness, because it would lack some of the moral or interest-based motivations that trusters characteristically ascribe to the objects of their trust. Finally, we have argued that in the case of socio-technical systems, which involve humans as operators, trust in the system's operators cannot be transferred straightforwardly to the system as a whole.

References

Baier, A. (1986). Trust and antitrust. *Ethics, 96*, 231–260.

Birolini, A. (2007). *Reliability engineering. Theory and practice*. New York: Springer.

Coleman, J. S. (1990). *Foundations of social theory*. Boston: Harvard University Press.

De Paoli, S., & Kerr, A. (2008). Conceptualising trust: A literature review, NIRSA working paper series No. 40, Available at http://eprints.nuim.ie/1145/ (accessed 10 August 2010).

Faulkner, P. (2007). On telling and trusting. *Mind, 116*, 875–902.

Franssen, M., & Jespersen, B. (2009). From nutcracking to assisted driving: Stratified instrumental systems and the modeling of complexity, paper presented at the *Engineering systems: Achievements and challenges*, Cambridge, MA, MIT. http://esd.mit.edu/symp09/submitted-papers/franssen-paper.pdf

Hardin, R. (2006). *Trust*. Polity.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72*, 63–76.

Horsburgh, H. J. N. (1961). Trust and social objectives. *Ethics, 72*, 28–40.

Huyke, H. J. (2003). Technologies and the devaluation of what is near. *Techné, 6*, 57–70.

Ihde, D. (1990). *Technology and the life world*. Bloomington: Indiana University Press.

Jones, K. (1996). Trust as an affective attitude. *Ethics, 107*, 4–25.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society; studies in sociotechnical change* (pp. 225–258). Cambridge: MIT Press.

Latour, B. (2002). Morality and technology; the end of the means. *Theory, Culture & Society, 19*, 247–260.

McLeod, C. (2002). *Self-trust and reproductive autonomy*. Cambridge: MIT Press.

Nickel, P. (2007). Trust and obligation-ascription. *Ethical Theory and Moral Practice, 10*, 309–319.

Nickel, P. (2009). Trust, staking, and expectations. *Journal for the Theory of Social Behaviour, 39*, 345–362.

Nickel, P. (2011) Trust in technological systems. In M. J. de Vries, S. O. Hansson, A. W. M. Meijers (Eds.), *Norms and the artificial: Moral and non-moral norms in technology*. Springer, forthcoming.

Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines, 20*, 243–257.

Verbeek, P.-P. (2005). *What things do; Philosophical reflections on technology, agency, and design*. University Park: The Pennsylvania State University Press.

Verbeek, P.-P. (2008). Morality in design; design ethics and the morality of technological artifacts. In P. E. Vermaas, P. Kroes, A. Light, & S. A. Moore (Eds.), *Philosophy and design: From engineering to architecture*. New York: Springer.

Whittingham, R. B. (2004). *The blame machine: Why human error causes accidents*. Amsterdam: Elsevier Butterworth-Heinemann.

Winner, L. (1985). Do artifacts have politics? In D. Mackenzie, & J. Wajcman (Eds.), *The social shaping of technology*. Milton Keynes: Open University Press.

Winner, L. (1992 (1977)). *Autonomous technology; Technics-out-of-control as a theme in political thought*. Cambridge: MIT Press.