**Trust in engineering**

[Manuscript version. Final version to appear in D.P. Michelfelder & N. Doorn, eds., *Routledge Companion to Philosophy of Engineering* (Routledge, forthcoming).]

**Philip J. Nickel, Eindhoven University of Technology**

Abstract:
Engineers are traditionally regarded as trustworthy professionals who meet exacting standards. In this chapter I begin by explicating our trust relationship towards engineers, arguing that it is a *linear but indirect* relationship in which engineers "stand behind" the artifacts and technological systems that we rely on directly. The chapter goes on to explain how this relationship has become more complex as engineers have taken on two additional aims: the aim of social engineering to create and steer trust between people, and the aim of creating automated systems that take over human tasks and are meant to *invite* the trust of those who rely on and interact with them.

Keywords:
Trust, professionalism, social engineering, trust in automation

## I. Introduction

According to one account, after the (first) Quebec Bridge Disaster in 1907 "it was immediately recognized that a serious blow had been struck to public confidence in the whole engineering profession" (Roddis 1993, 1544). Responsibility for the Quebec Bridge Disaster is largely ascribed to the consulting engineer, Theodore Cooper. In his design, Cooper extended cantilevered bridge construction beyond its earlier scope of application without sufficient testing. He was out of touch with the on-site engineers and workers who observed the bridge show increasing signs of strain and material failure during construction. In the bridge failure dozens of workers died, and after the collapse of the structure the two sides of the Saint Lawrence remained unbridged near Quebec City until 1919 (except when the river was frozen over).

In Canada, this disaster and a second that followed it at the same location led to a new awareness of professional responsibilities of engineers, as embodied in the Ritual of the Calling of the Engineer (ibid., 1545). Even today, many newly certified Canadian engineers pledge awareness of their responsibility towards society and receive a special ring to wear as a reminder of this professional responsibility.[1] One plausible explanation of this practice is that it instills and expresses the value of trustworthiness within the engineering profession, in much the way that the Hippocratic Oath and its associated symbolism functions within the medical profession.

[1] "The Ritual of the Calling of an Engineer," University of Guelph Department of Engineering, https://www.uoguelph.ca/engineering/events/2018/03/ritual-calling-engineer (accessed 28 September 2018).

Trust and trustworthiness are complementary attitudes. Trust, roughly speaking, is the expectation of one person or entity, that a second person or entity will uphold their commitments and meet certain standards. Trustworthiness, on the other hand, is the disposition of a person or entity to perform in the way that others reasonably expect it to perform, given relevant commitments and standards. For engineering to be trustworthy means that those designated as professional engineers can be reasonably expected to carry out certain tasks such as bridge construction to a high standard. When these trust expectations are disappointed, as in the Quebec Bridge Disaster, the salience of trust and trustworthiness suddenly become obvious. In the words of Annette Baier, one of the founding figures in bringing philosophical attention to trust, "We inhabit a climate of trust as we inhabit an atmosphere and notice it as we notice air, only when it becomes scarce or polluted" (Baier 1986, 234).

In this chapter, my agenda is threefold. First, I consider the traditional notion of engineers as trustworthy professionals with particular competencies, in order to see what this implies for the ethical orientation of engineers. Second, I consider how engineers try to "engineer trust" in contexts where agents interact, so as to achieve a certain desired form of interaction. This form of social engineering raises epistemic and moral questions about whether trust within these designed contexts is correctly based on the reality of these contexts, and about the unintended consequences (such as "filter bubbles") that sometimes result. Third, I consider trust in engineered systems themselves, especially "smart" systems that collect data and use it to make automated decisions (or to advise humans how to do so) and take over human tasks. I argue that there are important epistemic and moral questions about what type and amount of evidence is needed when relying on such systems. In addition, philosophical questions are raised by the very fact that researchers are inclined to call such reliance on automation "trust."

Because of this manifold agenda, the notion of trust and trustworthiness that I consider in relation to engineering is multifaceted. I do not expect to develop an overall definition of trust covering all these cases.[2] It will be useful at various points to consider the explanatory work being done by the concept of trust as contrasted with other possible attitudes of reliance. It is especially useful to have on hand a contrasting notion of *strategic reliance*, in which one person or entity relies on a second person or entity on a purely pragmatic basis, without the first at any time thinking that the second *should* take this reliance into account, or that the second is committed to behaving in accordance with particular norms. When I drive over a bridge, what is the difference, after all, between saying that I *trust* it to hold the weight of my vehicle, and saying that I merely rely on it to hold that weight? Arguably, trust plays no distinctive explanatory role in such situations.[3] When we talk about trust as an attitude distinct from such strategic reliance, we have in mind something affectively and

[2] How trust is theorized depends on one's explanatory agenda. For example, in behavioral economics and political theory, trust is often theorized as a mechanism through which cooperation emerges from interactions between individuals or state actors. In the theory of child development, by contrast, trust is theorized as a normal psychological disposition of the child to rely comfortably on the primary carer(s). I borrow elements from these diverse conceptions of trust as the context demands. For philosophical approaches to the concept of trust, see Nickel 2017, Simpson 2012, McLeod 2002.

[3] But see Nickel 2013b.

morally loaded. Such an attitude does distinctive explanatory work in accounting for cooperation and the nature of relationships between individuals.

## II. Trust in engineers as professionals: linear but indirect

The first thing we often think of when we consider trust in engineering, is trust toward engineers as professionals. Engineers take on special responsibilities toward clients, colleagues, technology users, and society, in virtue of representing themselves as professionals. (See the chapter “Responsibilities to the public” in this collection.) Engineers are held to have a responsibility of “non-maleficence” (non-harm) to society and to users of technology. More broadly, engineers are committed to carrying out work on the basis of up-to-date scientific and technical expertise, in line with best practices for the relevant engineering discipline. These responsibilities and stringent standards are explicitly stated in nearly every code of professional ethics of the various engineering disciplines (see Chapter [insert name of Davis chapter] for additional discussion). When taken at face value, such explicit commitments signal the trustworthiness of engineers, and provide a strong interpersonal basis for trust in engineering as a profession.

The sociology of professions contains trust as a significant (Evetts 2006, Brown & Calnan 2016) but limited (Adams 2015) theme. One of the main purposes of professional designations is to provide a readily-accessible reason for trusting those persons on which they are conferred. The professional designation works as follows: “Education, training and experience are fundamental requirements but once achieved (and sometimes licensed) … the exercise of discretion … based on competencies is central and deserving of special status. … Because of complexity it is often necessary to trust professionals’ intentions” (Evetts 2013, 785).

However, compared with the study of trust in, and within, other professional disciplines such as medicine and law (e.g., Brown & Calnan 2012, O’Neill 2002, Hall 2002), there has been very little empirical and ethical research on trust in professional engineering One reason for this is that engineers do not interact directly with members of the public the way that physicians, nurses, and lawyers do. In other professions, direct personal trust in known persons often remains essential to the delivery of service. By contrast, most people’s reliance on engineering is impersonal. Users interact with cars, bridges, and heart monitors, and unidentified engineers are presumed to have designed these technological artifacts with suitable care. It is a *linear* trust relationship (USER → TECHNOLOGY →ENGINEER), but an *indirect* one.

Despite the lack of scholarly study of trust in the engineering profession, we can still find important clues about how this *linear-but-indirect* trust relationship has become more complex over time. Emotional reactions such as blame toward engineers and others in the aftermath of accidents reveal the occasion and the object of distrust, and by extension the people and entities that one was inclined to trust in the first place. In its historical context, the Quebec Bridge Disaster caused distrust in Cooper and in engineers more broadly. However, in more recent engineering failures such as Dieselgate and the bridge failure in Genoa, the failure of a technology seems to have caused distrust in corporate entities (e.g., Volkswagen, Autostrade) and government oversight bodies, instead of or in addition to distrust in engineering as a discipline. This historical shift in the object(s) of distrust following cases of technological failure is partly due to a diffusion of responsibility. One important factor is the state's assumption of responsibility for technological risk. According to Beck's "risk society" theory, in the Twentieth Century, government entities assumed more responsibility for the overall balance of technological hazards, compared with the "limited state" of prewar Western modernity (Beck 1992). One scholar describes Beck's theory as follows: "With the beginning of societal attempts to control, and particularly with the idea of steering towards a future of predictable security, the consequences of risk become a political issue. … It is societal intervention — in the form of decision-making — that transforms incalculable hazards into calculable risks" (Elliott 2002, 295). When the state assumes responsibility for the unwanted effects of technological development, responsibility for technological risks, as well as trust and distrust, are diffused away from engineers (Renn 2008, 28).

Managerial changes of the last forty years have further diffused both responsibility for and trust in the engineered world around us. The advent of the "audit society" has placed some of the responsibility for technological risks with large companies and non-state governance bodies that rationalize technological risks from the perspective of management (Power 1997). Practices of *internal control* have also emerged in which the task of oversight is partly delegated by the government to organizations themselves and accountancy firms (Power 2007). These practices have made trust in engineering, which was already indirect and impersonal, more diffuse, complex, and abstract (see Figure 1). They may also lead to a paradoxical yearning for more personal forms of trust (Brown & Calnan 2012), and the need, mentioned by Evetts above, to "trust professionals' intentions" in situations of complexity (op cit.). Perhaps the yearning for personal trust in our impersonal engineered environment leads us to seek channels of focused expression for trust or distrust, for example in the "smart" technological products and services we use (e.g., Apple devices, Facebook) or in the celebrity technology entrepreneurs we associate with them (e.g., Steve Jobs, Mark Zuckerberg).

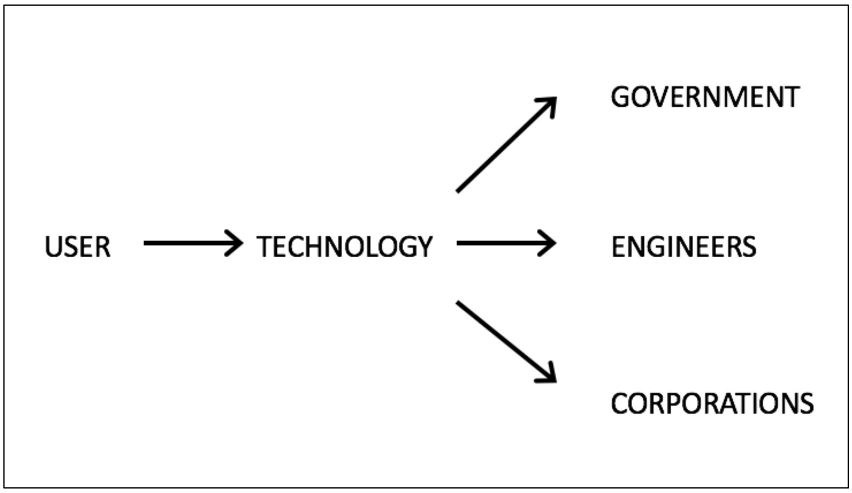

**Figure 1. Diffusion of linear-but-indirect trust in engineering**

Figure 1 suggests that the immediate object of trust in engineering is an *artifact or technology that we use practically* in some way. The linear-but-indirect relationship has become more complex because what stands behind our immediate relationship to an artifact or technological system is a whole array of people and institutions.

By way of contrast, it is interesting to compare trust in engineering with trust in science, where science is thought of as a body of knowledge and a way of interpreting experience. We might say that trust in science is not usually directly manifested in acts of *practical reliance*, but rather in acts of *believing in* certain established scientific claims, as well as adopting certain interpretations of experience. When scientific knowledge is crystallized into an act of concrete practical reliance (e.g., having a medical device installed, or trying out a new way of fertilizing the soil), we are inclined to redescribe it as trust in engineering. In this way, even though science and engineering are inherently related, trust in engineering remains conceptually independent of trust in science because it is (in philosophical terminology) practical rather than theoretical.

## III. Engineering interpersonal and interagential trust

There is widespread awareness that technology shapes human relations. Engineering sometimes has the explicit design goal to shape or increase trust relationships between people. Architecture, urban planning, web design, industrial engineering, and industrial design are all fields that can take this kind of "social engineering" as a goal. In what follows I focus on the design of digital environments as a way of illustrating a general trend toward reflexivity in the practice of (social) engineering that influences trust relationships. For example, in scholarly literature about online interactions starting about twenty years ago, concerns were voiced about the sustainability of trust relationships online, where the interacting parties are anonymous or pseudonymous, and not answerable for their words and deeds (Pettit 2004). Anonymity and the inability to rely on reputation in online environments have been used to explain fraud in one-off market exchanges in the digital sphere, as well as trolling behavior. Fraud within market exchanges has been empirically linked to the logic of one-off transactions between rationally self-interested agents with no reputation to protect (Ba, Whinston, & Zhang 2003). Concerns about trolling have also been empirically grounded in an "online disinhibition effect" in which people experience less inhibition to behave in impulsive or antisocial ways when online (Suler 2004).

These effects were originally unintended side effects rather than desired effects of participation in online platforms, but as they became increasingly well understood, they have been taken into account in engineering design (Resnick & Zeckhauser 2002). Indeed, since these effects are predictable and empirically grounded, engineers of networked environments have a responsibility to take them into account (Friedman, Kahn, & Howe 2000). Similarly, those who design environments for offline interaction, such as architects, urban planners, and industrial engineers, must keep in mind similarly relevant social scientific knowledge about the predictable effects of their interventions.

Recently, other effects on trust in digital environments and social media have been studied, related to the digital transformation of society itself. For example, trust and distrust have been linked to pervasive personalization and network effects. Turcotte et al. (2015) found that recommendations of news items by Facebook “friends” perceived as “opinion leaders” increase trust in the source of these items. One can easily imagine that with billions of people reading news primarily on or via social media, such trust phenomena could strongly influence people’s worldviews and political beliefs, leading to phenomena such as “fake news”.

In design practice, the implications of such findings depend on many additional premises about the ecology of social media. For example, if opinion leaders tend to have larger networks than other users, then they should have outsized effects on trust in news sources. Furthermore, if people tend not to have *contradictory* opinion leaders within their social networks, these trust effects would be likely to be one-sided, leading most individuals to trust only a select (and potentially biased) set of sources. It is tempting to explain political polarization partly in terms of such effects. Recently, the dark side of social engineering of social media has captured public attention, associated with concepts such as “filter bubbles” and “fake news”.  Social media has been held accountable for manipulation. This type of explanation may be elusive, however, because the social media ecology is constantly changing with technological, economic, and social forces. When digital natives who have had no experience of political and social life before social media become the norm, emergent phenomena observed within social media may change fundamentally, and will also be evaluated differently (for example, because norms of privacy have evolved).

As a result, the effects of design on social interactions, as well as how these interactions are experienced and evaluated by participants, are not easily predictable. They depend to a large extent on conventions, culture, and user practices. In order to anticipate and plan for these effects, it is useful to combine empirical, technical, and ethical inquiry during the design process in order to achieve the right balance of trust within designed social environments. Such combined inquiry has come to be called “value sensitive design,” “design for values,” or “responsible innovation,” and has been applied explicitly to the value of trust on a number of occasions (Friedman, Kahn, & Howe 2000; see examples in Nickel 2014). See the chapter [insert Chapter 27 title?] for a broader discussion of such methods.

Innovation that takes the value of trust into account must face the complexity of the concept and its value. Trust has two dimensions of value: practical and epistemic. On the practical dimension, trust is constitutive of healthy relationships. Trusting and being trusted have an intrinsic practical value for friends, neighbors, and co-citizens. They also have instrumental value in the sense that they help to obtain the benefits of cooperation and compliance. On the epistemic dimension, the value of trust is determined by whether it is well-grounded. One person should trust a second person just to the extent that that second person really is trustworthy or reliable: there should not be over- or under-trust. Furthermore, in some contexts it is important for the first person to have sound *reasons* for believing the second person is reliable, because not having such reasons would be irresponsible or reckless (Manson & O’Neill 2007; Voerman & Nickel 2017). This is particularly relevant when engineers are designing for other professionals, who have an obligation to act reasonably when trusting others. (The reasons at stake are often implicit, as van Baalen & Carusi (2017) argue in the case of multidisciplinary clinical decision-making.)

### IV. Trust in automation

The previous sections looked at two important kinds of trust in relation to engineering practice: first, trust in engineers themselves; and second, trust in other people whom we encounter within engineered environments. In this section I turn to a third kind of trust in engineering: trust in automated technologies. These technologies are often designed to invite trust, and the language of trust is used to talk about how we rely on them. In this section I will assess whether this talk of trust is superficial or serious. If it is merely superficial, then we can replace it with talk of strategic reliance without any explanatory loss. I will argue that trust in automation should be taken seriously.

"Automation" here refers to technological systems that take over human tasks requiring some intelligence and skill. Examples include trust in a robot to take over parts of a shared assembly task, trust in artificial intelligence to make decisions in a shared medical task such as a surgery, but also more "casual" artificial agents such as those that assist in everyday decisions (e.g., about what music playlist to put on at a party).

A useful point of reference is a recent review article synthesizing a model of trust in automation on the basis of over a hundred selected empirical studies of the subject (Hoff & Bashir 2015). The model distinguishes two phases of trust: the "initial learned trust" that one develops prior to interacting with the automated system, and the "dynamic learned trust" that one develops subsequently during one's interaction with the automated system. The overall picture of trust and its determinants can be seen in Figure 2.

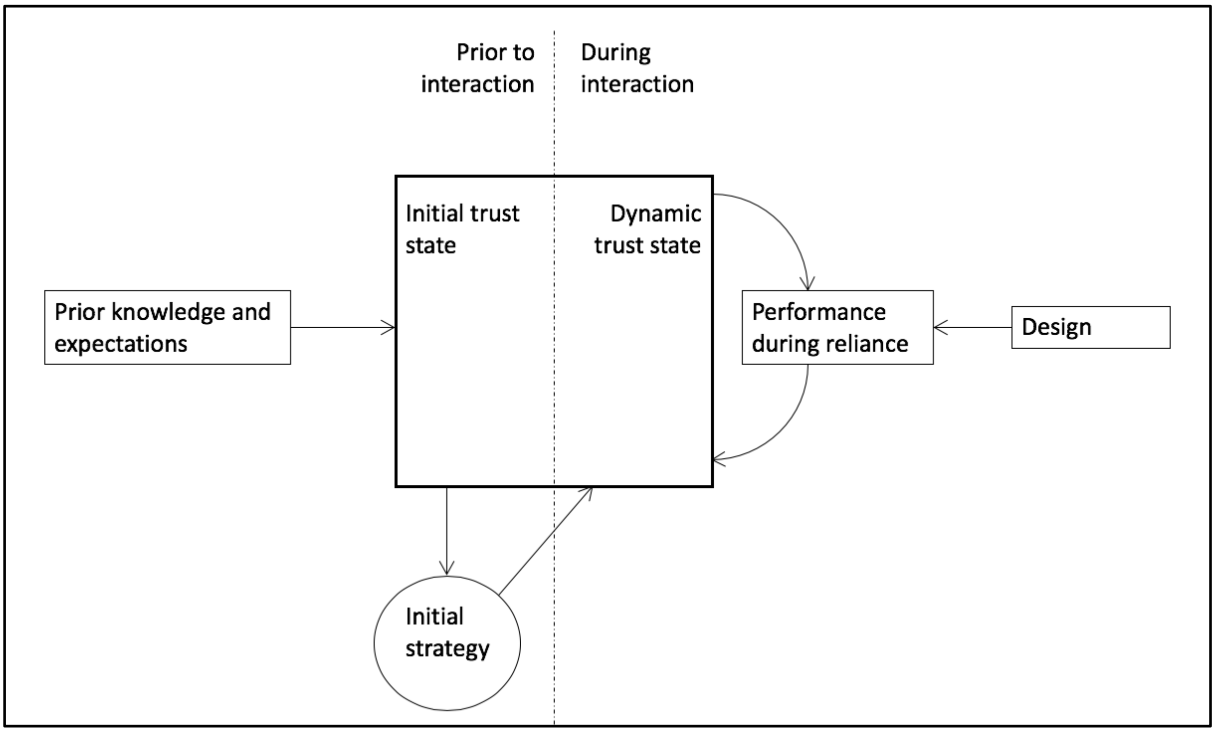

**Figure 2. Trust in automation (simplified version of figure from Hoff & Bashir 2015).**

Hoff & Bashir follow Lee & See's definition of trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (2004, 54). This definition involves a belief (or related representation) that automation will have *instrumental utility* within a context of uncertainty and vulnerability.

The trouble with this definition, and the scholarship from which it is derived, is that we do not know whether to take the talk of trust seriously. The literature does not clearly distinguish between trust and strategic reliance, defined earlier as calculative, instrumental reliance. When I rely strategically on somebody or something, this does not carry the implication that the relied-upon entity *should* take my reliance into account, nor the thought that the relied-upon entity *is committed to* or *is responsible for* behaving in accordance with any particular norms.

We must therefore ask, *what is gained by thinking of reliance on automation in terms of trust*? Answering this question is a serious challenge because it requires us to consider what we expect from automation itself; such that it, unlike ordinary artifacts such as bannisters and buses, might be the object of genuine trust. One possible answer is that there is nothing to be gained. Scientists are simply using the word “trust” in an extended and casual sense. They have no interest in a distinctively normative notion of trust, nor in an explanatory contrast between trust and strategic reliance.

This answer is unsatisfying. Automation has been introduced in areas of life such as medicine that are full of ethically weighty decisions. Inviting trust in that automation carries significant ethical consequences. It is doubtful that the word “trust” has been carelessly chosen; it seems to invite rather than avoid these ethical implications. For the sake of charitable understanding, we must at least try to find a different answer to our question.

Another possible answer is that trust in automation is a distinctive normative attitude, but an indirect one. We saw earlier that trust in engineering is linear but indirect. The engineer is an indirect object of the attitude of trust, standing behind the technological artifact with which users interact. In a stage play, the director and the author are off-stage, but they take most of the responsibility — and the criticism — for the performance. Analogously, one might maintain that the reason we talk about trust in automation is because engineers (or companies) stand in appropriate relations to that automation.

However, there are two problems with this account. The first is that automation itself is not a real object of trust according to this story. The engineer, alongside other institutional entities, is the object of genuine trust. But we do not actually rely on engineers for the particular “performances” that we require during the activity of carrying out the shared task. Hence their agency is insufficiently related to what we actually rely on to make it count as trust (Nickel 2013). The second problem is that this story holds for buses, bridges, and bannisters, just as well as for automated systems. The notion of trust at stake is therefore just as thin as for other everyday artifacts. In that case, there is nothing special about automation.

A third possible answer holds that the perceptible features of automation technology *invite* distinctive trust attitudes. There are several features of automation technology that make it psychologically a suitable object of trust. Some of these are discernible in the scientific literature. Hoff & Bashir summarize advice in the literature for “creating trustworthy automation”, such as: “increase the anthropomorphism of automation” taking into account “age, gender, culture, and personality”; use “gender, eye movements, normality of form, and chin shape of embodied computer agents to ensure an appearance of trustworthiness”;

and “increase politeness” (Hoff & Bashir 2015, 425). From these strategies it is clear that many automation technologies take on human social, bodily, and communicative characteristics that make them appear different from dumb artifacts or mere things in our environment. Anthropomorphic robots are among the most striking examples of automated systems that elicit interpersonal trust attitudes because of their human form and characteristics (Coeckelbergh 2012).

On its own, however, this third answer seems shallow. Although we might be happy to talk about trust in automation because the design of automation exploits human trust cues, on further reflection talk of trust should be seen as a useful but insignificant manner of speaking, at most an explanatory stance (Dennett 1989). Since we *know* that anthropomorphic characteristics that invite trust are design add-ons, and reflect no underlying personhood, they do not ultimately support robust talk about trust in automation. They may help us understand the phenomenology of the technology user’s reliance on automation (and the cunning of the designer!) but they do not reveal a genuine trust relationship. In a recent article, Tallant (2019) puts forward a view along these lines.

The final answer to be considered looks at the fact that automation technologies are complex sociotechnical systems with a high degree of intelligence and rationality built into them both functionally and contextually. Functionally, they are capable of perceiving and responding to the environment. We evaluate them not just as being either reliable or unreliable (like a bridge), but as having correct or incorrect representations, and as drawing correct or incorrect inferences. These functional aspects of sophisticated automation make them subject to some of the same kinds of normative evaluation that also underwrite our trust attitudes. For example, an artificial agent such as Watson that gives answers in response to questions invites many of the kinds of normative evaluation of its speech — whether it has made a relevant and sincere assertion, whether it has interpreted a question correctly, etc. — as a human speaker does. It is not at all surprising that we would describe our attitude toward such an agent in terms of trust (Nickel 2013a). We do not need to countenance automation as having free will or moral capacities in order for some of these normative evaluations to be appropriate. Such a notion of trust may be thinner than interpersonal trust, but it is not obviously thinner than the trust that we take towards institutions or the roles within them, for example.

Conversely, it is likely that automation technologies will also begin to assess the reliability of humans, or as it were, the quality of the working relationship between automation and human, when the functioning of the automation depends on humans doing their part within that working relationship. Whether we call such a faculty a capacity for *trust*, or not, will depend on both the specific character that the capacity takes, and the matter of what we can gain, explanatorily, from using trust concepts to describe it.

In sum, it is plausible that reliance on automation involves a distinctive attitude of trust, contrasted with strategic reliance. The three factors that support talking about complex automation in terms of trust — the indirect role of the human engineer, the designed social cues that invite trust in the automated systems themselves, and the sociotechnical complexity of these systems (including representational states and perceptual and inferential processes) — are mutually reinforcing of one another.

Suppose we grant that real trust in automation is possible in cases where the aforementioned factors are present to a high degree. There are still open questions concerning whether and when we *should* trust automation. Philosophically, we must take a critical epistemological and ethical view of trust practices. Trust is not a goal to be unreflectively increased or a purely psychological given. People have *reasons* for trusting, both contextual reasons (e.g., there are few better options) and broadly evidential reasons (e.g., the trusted entity is seen to have the trustor's interests at heart). These reasons can be evaluated for their epistemological and ethical adequacy.

Engineers can design automated systems to manipulate our trust. Suppose an artificial agent is made to mimic a user's facial appearance and voice in order to gain their trust (Verberne, Ham, & Midden 2015). Such a strategy for increasing trustworthiness counts as *deceptively manipulative* in circumstances where increases in perceived trustworthiness do not correspond with good epistemic reasons for trust or greater usability (Spahn 2012; Smids 2018).

A theory of justified trust in automation would be useful as a starting point for ethical and epistemological evaluation, but currently no account of it exists. At a minimum, we want two things from such an account. First, we want a criterion for over-trust and under-trust, something that tells us when trust is appropriate. Such a criterion or norm concerns the would-be trust*or*, the one who is put in a position to trust. Let us call this the "user norm". Second, we want an account that illuminates the obligations and responsibilities of the engineer, the designer of automation. We want, for example, to know what would count as wrongful manipulation and exploitation of trust. Let us call this the "design norm". (Simon's (2015) identification of "epistemic responsibility" and "governance/ design" as two relevant aims of an account of epistemic trust in the digital sphere is analogous to the distinction I draw here between user norms and design norms. Simon's account is not directly applicable to trust in automation because her account is explicitly tailored to epistemic trust involving belief or knowledge, rather than acts of practical reliance.)

To spur our imagination a bit, consider a not-too-improbable case of trust in automation (based on Google's service *Hire*). Suppose the human resources department of a company or university begins to use artificial intelligence to filter job applications and reduce the "grunt work" of recruitment officers who previously sifted through hundreds or thousands of files. Suppose that the recruiters care professionally about the values of fairness and diversity, and suppose that this AI tool has built-in features that are intended to increase fairness in recruitment, such as ignoring sex, age, and name when identifying salient candidates. Under what conditions should a recruiter trust such automation to suggest salient candidates?

I propose two main conditions for well-grounded trust of a user U in an automated agent A. Because of the lack of scholarly literature on the topic, these conditions are meant as hypotheses for discussion rather than a battle-tested theory. The first condition is meant to track the idea that the automation does not operate on reasons that are completely irrelevant or contradictory to the reasons that the user has (or should have). For example, is ignoring demographic characteristics sufficient to secure the values of fairness and diversity in recruitment? The recruitment officer might reasonably hold that affirmative action is required to secure these values, which would imply being *aware* of demographic characteristics rather than ignoring them. In that case, trusting this form of automation might lead to a sort of betrayal of the recruiter's values.

The second condition is meant to track the idea that the user must have sufficient reason to believe that the actual performance of the automation is or will satisfy standards of competence at carrying out the tasks at hand while not undermining U's interests. This sufficient reason can come from various sources depending on whether, as in Figure 2, the user is in an "initial state" prior to actually interacting with the automation technology, or is in a "dynamic state" in which there is some experience with the performance of the system. For example, before relying on the automation technology, a particular recruiter might hear from a trusted colleague that the system has improved the quality and diversity of candidates, while freeing up valuable time for recruiters to communicate quickly with those candidates.

Accordingly, let us then hypothesize the following conditions on warranted trust in automation:

> User U's trust in automated agent A to do task T is reflectively warranted if and only if the following two conditions are met:
>
> (**Hypothetical Transparency Condition**)
> If U knew and rationally reflected upon the reasons that influenced A in relation to T, U would recognize the relevance and strength of these reasons from the perspective of U's own interests;
>
> (**Internal Adequacy Condition**)
> U has either experienced A doing T adequately or has other relevant support for her expectation that A can T adequately, such as the judgment of epistemic peers to the effect that A can T adequately, or the normative commitment of those who designed and deployed A to take U's interests into account when doing T.

Annette Baier proposes a similar hypothetical transparency test of the moral adequacy of trust (Baier 1986). However, she rejects anything like the internal adequacy condition as a general criterion for the moral adequacy of trust, because she points out that children are often perfectly warranted in their trust even though they do not have any such reasons (ibid.). By contrast, I do adopt such a condition here because *reflective warrant* is meant to be appropriate for sovereign adults relying on automation. Particularly for professionals like those in our example who are in a position to trust automation in the workplace, it is plausible that they need to have some kind of positive reason for doing so. The idea of reflective warrant is meant to capture this idea. When children rely on automation, a different standard applies.

When coupled with the claim that users often have a strong (moral) interest in having a warrant for trust in automation, it follows that engineers who design automation have an obligation, other things equal, not to deceive people about the grounds for trust, and that they may have a positive obligation to give users access to information that helps fulfill the conditions of the warrant. In this way, we derive a relevant “design norm” against deception and exploitation from the “user norm”. This should remind us of the point we started with in this chapter, that there is always a connection between trust and trustworthiness, between the user’s trust and the engineer’s trustworthiness. It is not a shallow connection, but a deep and ethically significant one.

Such an account of warranted user trust in automation also carries implications for governance, suggesting a new interest at stake in Europe’s General Data Protection Regulation (GDPR). That law asserts that individuals have a right to obtain “meaningful information about the logic involved” that is used in automatically processing their personal data (Regulation 2016/679, 2016, 2.15.1.h), and data controllers have a corresponding duty to provide such information (ibid, 2.13.2.f). Most people have understood the aim of the GDPR as protecting one’s interest in confidentiality and control over one’s personal data. But there is in fact a second interest at stake here: trust in automation. When personal data processing is being carried out in order to provide automated services (whether at work or at home), the user has an interest in “meaningful information about the logic involved” *in order to safeguard the Hypothetical Transparency Condition*. This would include information about *the reasons for* using data inputs of a certain type and the *design thinking* behind the automation: reasons that help explain why the automation behaves as it does. These reasons should normally be conveyed during the mundane process of training a person to work with an automated system.

## V. Conclusion

The opening question of the chapter can now be examined from a new perspective: what will trust in the engineer and the engineering profession look like in the future? Engineers may need to rethink how they want to be trusted, and to regard not just reliability and trustworthiness, but *trust* as an important commodity in itself. Brown & Calnan (2012) and Evetts (2013) emphasize the way in which trust is used by people as a way of navigating institutional and organizational complexity. Trust will be in ever more demand in a future technological environment that is ever more complex and abstract. With engineers at the center of some of the most important technological changes in our future, the question of trust will likely arise for them in a pointed way.

Another way in which the future of engineering will not be similar to its past is that engineering will apply automation to itself. Calculative and mechanical tasks that used to form the core competences of engineering will be carried out by computers and robots. Engineers will design and care for socio-technical systems including both human and mechanical components using a broad set of skills: computer programming and data science, interdisciplinary communication, lifelong learning, ethical awareness, and critical design thinking. Designing for interpersonal trust, and for trust in automation, will be core competences. So although trust and engineering have many strands — as exhibited in the preceding sections — these look increasingly as if they will be woven together in the future of engineering. (For more on this topic see the chapter “Reimagining the future of engineering” in this volume.)

**Suggestions for Further Reading**

Classics in the philosophical literature on trust include Annette Baier's paper "Trust and Antitrust" (1986) and Richard Holton's "Deciding to Trust, Coming to Believe" (1994). For contributions to the recent philosophical literature on trust, see Faulkner & Simpson, eds., *The Philosophy of Trust* (Oxford, 2017). For an earlier treatment of challenges to the idea of trust in technology and the relationship of trustworthiness to reliability in engineering, see Nickel, Franssen, & Kroes (2010).

**Related Topics**

Reimagining the future of engineering
Responsibilities to the public
[Chapter 27: RRI/ Design for values]
[Davis chapter]

**References**

Adams, T.L. 2015. Sociology of professions: international divergence and research directions. *Work, Employment, and Society* 29: 154-165.

Ba, S., Whinston, A.B., & Zhang, H. 2003. Building trust in online auction markets through an economic incentive mechanism. *Decision Support Systems* 35, 3: 273-286.

van Baalen, S., & Carusi, A. 2017. Implicit trust in clinical decision-making by multidisciplinary teams. *Synthese*. DOI 10.1007/s11229-017-1475-z

Baier, A. 1986. Trust and antitrust. *Ethics* 96: 231-260.

Beck, U. 1992. *Risk Society — Towards a New Modernity*. London: Sage.

Bozdag, E., & van den Hoven, J. 2015. Breaking the filter bubble: democracy and design. *Ethics and Information Technology* 17, 4: 249-265.

Brown, P., & Calnan, M. 2012. *Trusting on the Edge: Managing Uncertainty and Vulnerability in the Midst of Serious Mental Health Problems*. Bristol: The Policy Press.

Brown, P., & Calnan, M. 2016. Professionalism, trust, and cooperation. In Dent, M., Bourgeault, I.L., Denis, J.-L., & Kuhlmann, eds., *The Routledge Companion to the Professions and Professionalism*. Routledge. Pp. 129-143.

van Burken, C. 2014. *Moral Decision Making in Network Enabled Operations.* Doctoral Dissertation. Simon Stevin Series in the Ethics of Technology.

Coeckelbergh, M. 2012. Can we trust robots? *Ethics and Information Technology* 14: 53-60.

Dennett, D.C. 1989. *The Intentional Stance*. Cambridge, MA: MIT Press.

Elliott, A. 2002. Beck's sociology of risk: a critical assessment. *Sociology* 36: 293-315.

Evetts, J. 2006. Trust and professionalism: challenges and occupational changes. *Current Sociology* 54: 515-531.

Evetts, J. 2013. Professionalism: value and ideology. *Current Sociology Review* 61: 778–796.

Friedman, B., Kahn, P.H. Jr., Howe, D.C. 2000. Trust online. *Communications of the ACM* 43: 34–40.

Hall, M.A. 2002. Law, medicine, and trust. *Stanford Law Review* 55: 463–527.

Hoff, K.A., & Bashir, M. 2015. Trust in automation: integrating empirical evidence on factors that influence trust. *Human Factors* 57: 407–434.

Holton, R. 1994. Deciding to trust, coming to believe. *Australasian Journal of Philosophy* 72: 63-76.

Lee, J.D. & See, K.A. 2004. Trust in automation: designing for appropriate reliance. *Human Factors* 46: 50–80.

Manson, N.C. & O'Neill, O. 2007. Rethinking Informed Consent in Bioethics. Cambridge University Press.

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

Nickel, P.J. 2013a. Artificial speech and its authors. *Minds and Machines* 23, 4: 489–502.

—. 2013b. Trust in technological systems. In M.J. de Vries, S.O. Hansson, and A.W.M. Meijers, eds. *Norms in technology: Philosophy of Engineering and Technology, Vol. 9* Springer: 223–237.

—. 2014. Design for the value of trust. In J. van den Hoven, I. van de Poel, P. Vermaas, eds., *Handbook of Ethics, Values and Technological Design.* Berlin/ Heidelberg: Springer-Verlag.

—. 2017. Being pragmatic about trust. In P. Faulkner and T. Simpson, eds., *The Philosophy of Trust*. Oxford University Press: 195–213.

Nickel, P.J., Franssen, M., & Kroes, P. 2010. Can we make sense of the notion of trustworthy technology? *Knowledge, Technology and Policy* 23: 429–444.

O'Neill, O. 2002. *Autonomy and Trust in Bioethics*. Cambridge University Press.

Pettit, P. 2004. Trust, reliance and the internet. *Analyse & Kritik* 26: 108–121.

Power, M. 1997. *The Audit Society: Rituals of Verification*. Oxford University Press.

—. 2007. *Organized Uncertainty.* Oxford University Press.

Regulation (EU) 2016/679 of the European Parliament and of the Council. 2016. *Official Journal of the European Union.* L 119/1-L199/88

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

Resnick, P., & Zeckhauser, R. 2002. Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system. in Michael R. Baye (ed.) *The Economics of the Internet and E-commerce (Advances in Applied Microeconomics, Volume 11).* Emerald Group Publishing Limited. 127–157.

Roddis, K. 1993. Structural failures and engineering ethics. *Journal of Structural Engineering* 119: 1539–1555.

Simon J. 2015. Distributed Epistemic Responsibility in a Hyperconnected Era. In: Floridi, L., ed. *The Onlife Manifesto*. Springer. DOI 10.1007/978-3-319-04093-6_17

Simpson, T. W. 2012. What is Trust? *Pacific Philosophical Quarterly* 93: 550–569.

Smids, J. (2018). *Persuasive technology, allocation of control, and mobility: an ethical analysis.* Doctoral dissertation. Eindhoven: Technische Universiteit Eindhoven.

Spahn, A. 2012. And lead us (not) into persuasion… Persuasive technology and the ethics of communication. *Science and Engineering Ethics* 18: 633-650.

Suler, J. R. 2004. The online disinhibition effect. *CyberPsychology and Behaviour* 7, 321–326.

Tallant, J. 2019. You *can* trust the ladder, but you shouldn't. *Theoria.* DOI: 10.1111/theo.12177

Verberne, F.M.F., Ham, J., & Midden, C.J.H. 2015. Trusting a virtual driver that looks, acts, and thinks like you. *Human Factors* 57: 895–909.

Voerman, S.A. & Nickel, P.J. 2017. Sound trust and the ethics of telecare. *Journal of Medicine and Philosophy* 42: 33–49.