Chapter 10
Trusted Autonomy Under Uncertainty

Michael Smithson

10.1 Trust and Uncertainty

10.1.1 What Is Trust?

The main goal of this chapter is to elaborate the connections between trust, distrust, and uncertainty. Trust will be treated primarily as a psychological state, but sometimes also as a type of relationship or in purely behavioral terms. Trust, after all, is a social concept. However, it is not a contractual relationship, and so "trust" here will not have the kind of meaning in legal or institutional forms such as "trust fund" or "company trusts". The main theme is that trust and distrust inherently involve uncertainty (and risk) in two respects. First, uncertainty is a given in trust or distrust as a psychological state. Second, the processes in a trust or distrust relationship generate unknowns.

It may seem odd to begin this chapter by reconsidering definitions of trust and distrust, but this is necessary for three reasons. First, the concept of trust has been diversely defined in various disciplines, chiefly economics, psychology, political science, and sociology. These definitions often disagree with or talk past one another (see the discussion in [30]). Second, "trust", "mistrust", and "distrust" are multifarious in many natural languages. Each can be a noun or a verb, and each can describe a psychological state, a belief, a feeling, a relationship, or even (in the case of trust) a legal entity. And third, the terms have not been used consistently in the literature on human-robot interaction (HRI, from here on). Some HRI researchers have treated "trust" as synonymous with reliability, while others have brought in matters of trustee interest or intent regarding the trustor. "Distrust", on the other hand, has been relatively neglected in the HRI literature. Starting with "trust", we shall avoid the definitions used by some researchers that define trust by the ways in which it is

M. Smithson (✉)
Research School of Psychology, The Australian National University,
Bldg 39, Room 215, Canberra, Australia
e-mail: Michael.Smithson@anu.edu.au

H. A. Abbass et al. (eds.), *Foundations of Trusted Autonomy*, Studies in Systems,
Decision and Control 117, https://doi.org/10.1007/978-3-319-64816-3_10

formed [30]. Thus, while trust may arise from a rational choice (e.g., [16]), as a personality trait (e.g., [33]), or as an institutionalized or identity-based norm (e.g., [29]), none of these actually defines trust. Trust is defined in this chapter as an outcome of such choices, personalities, or normative processes.

A version of Hardin's [16] tripartite conceptualization of trust will be used here. In his framework, trust is defined in terms of attributes of the trustor, properties of the trustee, and the specific context in which trust occurs. A trustor may be "trusting" in the sense of an expectation (e.g., [18]), a positive feeling (e.g., [11]), or an attitude that has an intentional component (i.e., being willing to trust). According to a survey of 65 sources of definitions of trust [30], the typical characteristics ascribed to a trustee include predictability, reliability, competence, benevolence (toward the trustor), and integrity. Thus, for instance, a trustor may expect or feel that a trustee is reliable or benevolent.

Context includes the following components:

1. Dependence: What the trustor depends on the trustee to provide or to do,
2. Trust behavior: What the trustee must do to show and bestow trust,
3. Basis: Factors involved in the formation of trust, and
4. Stakes: The potential benefits and costs of trust.

The nature of the dependence involved in a trust relationship will also strongly influence the kinds of uncertainties involved, as will be elaborated later on. Trust behavior, as we will see later, involves a combination of deference to the trustee, relinquishment of control over or micro-management of the trustee, and relevant risk-taking. The basis for forming a trust relationship may be rational calculation, personal disposition, reputational, social identity, part of a role, or even part of a set of rules in an institutional setting [23]. The stakes can be financial or tangible, but also may include intangibles such as esteem, reputation, and even willingness to trust in future relationships. Moreover, the stakes may not be limited to direct consequences of trust, but also can include "side-effects" such as sociability, opportunity, and transaction costs and benefits.

What are the opposites of trust? The absence of trust, in the sense of indifference, clearly is not the same as distrust, mistrust, or paranoia. "Distrust" and "mistrust" often are used interchangeably, although common usage tends to construe "mistrust" in terms of suspicion or doubt about a target, and "distrust" as without doubt that the target is untrustworthy. A systematic treatment of distinctions between these two terms is beyond our scope, and the focus in this chapter will be on distrust. Several scholars have claimed distrust is the opposite of trust, whereby they mean an expectation or suspicion that the distrusted party is unreliable and/or malevolent (see [34]; or [12]: "I trust my friends; distrust my enemies"). Like trust, distrust is a social entity through and through. We can employ Hardin's tripartite framework for dealing with distrust, in a similar manner to trust, by considering it in terms of attributes of the distrustor, properties of the distrustee, and the specific context in which distrust occurs.

Distrust, then, incorporates attributions of unreliability and intentions toward the distrustor ranging from neglectful to malign. Distrustful relations therefore will be

characterized by hyper-vigilance, attempts to free oneself from any dependence on the distrusted party, and/or attempts to assert control over that party. Legally binding and enforceable contracts are an example of a relationship that could be based on distrust.

10.1.2 Trust and Distrust in HRI

How have trust and distrust been construed in the literature on HRI? Which definitions or conceptions are most useful in understanding HRI and designing technologies to implement or augment it? For instance, is trusting an AI-driven robot more like trusting a refrigerator, a trading bank, a surgeon, or a friend-or is it like none of these? Some scholars, such as Lee and See [26], have defined trust in HRI settings as rather similar to trust in humans. Lee and See's aspects of trust include performance, process, and purpose. The first two are similar to the well-worn concepts of reliability and predictability. The third refers to a belief that the automaton is functioning as its designers intended, and includes agreement with those intentions.

However, others have suggested that trust in HRI is not the same as human-to-human trust [28]. Jian et al. [20] found that people are more willing to rate an automaton than a human being as "distrusted", suggesting that there may be differences between human trust in automatons and human trust in humans. One source of such differences is that people tend to regard expert systems, AI systems, or computer-based decision support systems as more objective and rational than their human counterparts [7]. One could also add that people may expect automatons to have greater integrity than humans because they believe that automatons are not programmed to deceive. Of course, this stereotype could change rapidly as AIs become more sophisticated. In the HRI literature, the prospect of deception by robots (or AI) already has been raised [15].

There also is some evidence that people react more strongly to errors made by automatons than those made by humans, so that there is a swifter decline in trust (see [9, 43]). In a general sense, then, people may be less tolerant of uncertainties manifested in automaton behavior than in the behavior of their human counterparts. Humans, on the other hand, are expected to be more adaptive and creative than automatons, so it is plausible that novel actions or proposals from humans will be more trusted than if they come from automatons.

What is meant by "appropriate" trust in automatons? Oleson et al. [31] claim that appropriate trust of a system manifests itself in appropriate reliance on that system. Too much trust results in overreliance, and too little in insufficient reliance. However, it is worth bearing in mind that other factors can result in over- or under-usage of a system, such as a desire to avoid blame for bad outcomes (over-use) or a desire to gain credit for good outcomes (under-use).

Addressing the question of appropriate trust, Ososky et al. [32] refer to humans' tendency to anthropomorphize robots and to apply "inaccurate" mental models to inferences about robots' behaviours. Their remedy is that operatives have a full under-

standing of the automaton's capabilities and limitations. However, they do not systematically investigate the practical achievability of this suggestion. There already is an abundance of software and automated systems whose complexities exceed human capacity for anything approaching a complete understanding of their capabilities and limitations. Hancock et al. [15] recommend "transparency" in the form of system designs that are accessible and clear to human team members. However, there is an obvious potential for difficult tradeoffs or even dilemmas if one of the design objectives for a robot or AI also is that it is able to deceive enemies or even allies who are not cleared to know about that robot or AI.

Interestingly, the question of whether humans perceive (or can perceive) that they are trusted or distrusted by automatons seems to have been relatively neglected. At first glance, the question might seem nonsensical; surely we are not about to deem an automaton as being capable of trust. Nevertheless, the question makes sense for three reasons.

First, humans do anthropomorphize machines, so we cannot rule out the possibility that people may attribute an automaton's behavior towards them to trust or distrust of themselves by the automaton. This attribution certainly could arise when humans adopt what Dennett [6] called the "intentional stance". Dennett contrasts this stance against the "design stance", whereby an automaton's behavior is explained via beliefs about what it was designed to do. The intentional stance accounts for an automaton's behavior by assuming that it is a quasi-rational agent, with beliefs and desires of its own and the intelligence to pursue those desires on the basis of its beliefs. Moreover, people may be more likely to attribute trust to automatons than they would attribute emotions such as desire, because they are more willing to attribute belief states to automatons than feelings (cf. [19]) and the primary basis for trust is a set of beliefs.

Second, in connection with Lee and See's concept of purpose as a basis for trust, the intended uses of an automaton can include (dis)trust-relevant purposes such as monitoring its human teammates or deferring decisions to them. Thus, humans interacting with an automaton may adopt an intentional stance with regard to the automaton's designers and/or operators, attributing trust or distrust to these "puppeteers", even if they maintain only a design stance regarding the automaton itself. The automaton then manifests trust or distrust indirectly, via its apparently designed purposes and uses.

Third, Dennett's distinction between the design and intentional stances points to a candidate criterion for appropriate trust. A design stance would be appropriate in HRI most of the time. Automatons are indeed designed entities or systems; their designers will have had purposes and uses in mind. Trust based on a design stance will be limited to attributions of dependability, reliability, adherence to purpose-directed behavior, and the like. This kind of trust will be inappropriate only if the trustor has badly estimated the automaton's reliability or has misconstrued its design purposes. On the other hand, basing trust on an intentional stance clearly has pitfalls in the form of attributing benign intentions to an automaton. So, trust based on an intentional stance is likely to be inappropriate. However, as Dennett systematically argues throughout his book, the intentional stance works very well for predicting machine behavior, even when applied to something as simple as a thermostat. An

intentional stance is, as Dennett points out [6], a viable alternative when a design stance is not practical. This stance therefore is seductive and difficult to falsify.

Finally, imputation of trust to automatons also is important because, as automatons are made increasingly human-like, humans will interact with them in more social rule-following ways. Reciprocity is a key social rule governing many aspects of human-human interactions, and it is likely to become increasingly relevant to HRI. Trust and distrust often are reciprocated, so we may expect that people are more likely to (dis)trust an automaton if they believe that the automaton (dis)trusts them. In short, a trust-enhancing way of humanizing automatons is to enable them to manifest trust-like behaviors toward their human teammates. Whether or when this would be desirable is a matter for careful consideration by designers.

How should we measure or evaluate trust in human-automaton relations? Examples from the literature include the Human Computer Trust Rating Scale [25]. Yagoda and Gillan [44] propose a scale that taps four closely-related adjectives for describing different aspects of HRI: reliability, dependability, accessibility, and timeliness or predictability. An in-depth critical review of the relevant measurement issues is not within the scope of this chapter, but suffice it to say that measuring trust in HRI is an active area of research and the current state of the art is at a fairly preliminary level. It seems unlikely that a single scale or battery of scales will be adequate for all types and contexts of HRI, and that as AI and related technologies advance, the measurement of trust in automatons will need revising.

10.2 Trust and Uncertainty

10.2.1 Trust and Distrust Entail Unknowns

Trust as a psychological state entails willingness to take risks by placing oneself in a vulnerable position with respect to the trustee (e.g., [11, 16, 23]). Uncertainty is therefore a given in trust. Moreover, trust relations may have to be forged in contexts bearing unknowns. This would be the case with new complex technology, for example, even if it has undergone extensive testing.

The key connection between trust and uncertainty is that to enter a trust relationship requires at least some non-surveillance of the trustee, and at least temporary non-accountability (freedom from micro-management) for the trustee. Thus, the trustorforgoes an entitlement to place the trustee under 24–7 surveillance or total accountability. Thus, trust relationships create unknowns and require that the trustor tolerate them [37]. Relinquishment of knowledge and control is primarily what distinguishes trust relationships from contracts (or assurance). In effect, such relinquishment amounts to trading one source of uncertainty for another, in the sense that uncertainty about whether desired goals or outcomes will be attained is reduced via the trust relationship, which in turn imposes a less aversive kind of uncertainty regarding the means by which goals or outcomes will be reached, through allowing

the trustee discretionary power. This trade must be viewed by the trustor as worth enough to bear the risks entailed in a trust relationship.

Distrust as a psychological state amounts to a disposition to avoid being vulnerable to the distrusted party, often arising as a result of uncertainty about this party's intentions or future actions. Distrust therefore may involve unknowns in the form of suspicions as a given, or even as a justification for distrust in the first instance. Distrust also brings with it two additional forms of uncertainty. First, one may believe or suspect that the distrusted party lacks integrity and therefore doubt the veracity of information provided by that party. Second, distrust can morally license the manufacture of unknowns by the distrustor, either by withholding information from or outright deceiving the distrusted party. It seems plausible that people would find it easier to justify either of these acts if the distrusted party is an automaton than if it is a human, and therefore would be more likely to try to keep secrets from or deceive an automaton. The consequences of distrust in HRI appear to be relatively neglected in the research literature.

Relevant uncertainties can enter into any of Hardin's tripartite components: the (dis)trustor, the (dis)trustee, and the context. People can be unsure about their own psychological states; they may not be familiar with the automaton's reliability or design specifications; and they may have to engage in HRI in situations fraught with unknowns. Disentangling all of these uncertainties in a way that is relevant to trust considerations requires, first, ascertaining what is at stake in a HRI trust relationship. Thereafter, we can bring in knowledge about how and when people are likely to be able to tolerate and work with unknowns.

10.2.2 What Is Being Trusted; What Is Uncertain?

The relevance of uncertainties and their effects on trust in HRI will hinge on what is at stake in trusting an automaton. The stakes may be considered in terms of three aspects: The scope of the automaton's capabilities and responsibilities, the nature and sources of potential malfunctions or mishaps, and the kinds of errors or malfunctions committed by the automaton. The greater the scope of capabilities and responsibilities attributed to the automaton, the greater the impact of uncertainties about its functioning and capabilities on its trustworthiness. Likewise, greater perceived control over an important decision will be likely to increase the impact of uncertainties on trust. Yagoda and Gillan [44] present a useful two-dimensional framework regarding automaton capabilities. One dimension is the degree of intelligence and the other is the level of autonomy. AI would be high on both dimensions, while expert systems are typically high on intelligence but low on autonomy. A battery exemplifies low-intelligence but high-autonomy, and a robotic arm typifies low-intelligence and low-autonomy. It is plausible that being higher on either of these dimensions will increase the impact of uncertainties on trust in an automaton.

Turning now to the nature and sources of malfunctions or mishaps, two considerations are important to bear in mind. First, what kinds of errors or malfunctions

are most problematic? Suppose an automaton has a diagnostic function that makes a binary decision to raise an alarm or not. False alarms will be regarded as more harmful than misses in some settings (e.g., in a legal trial where false convictions are worse than false acquittals) but the reverse will be the case in others (e.g., diagnosing a contagious fatal disease, where false positives are not as harmful as false negatives).

Second, are the sources of potential malfunctions internal or external to the automaton? Trusting an automaton to function properly is one thing if the only possible causes of malfunction are hardware or software faults in the automaton itself. It is quite another if malfunctions could be caused by damage from attacks, sabotage, hacking, or other security breaches. This latter set of possibilities brings with it questions of trust regarding the automaton's robustness and security provisions, which may have little or no connection with its primary purposes or functions. Uncertainty about autonomy itself may raise doubts and concerns about who or what is controlling the automaton (e.g., whether it has been hijacked).

Finally, we turn to considering errors and malfunctions. Errors or malfunctions will break trust, although at least one study has suggested that they may not influence decisions of whether to permit the automaton to act [35]. This finding highlights the importance of separating considerations about trust from those regarding whether humans will override an automaton. The connection between these two matters is relatively unexplored. Reasons or explanations for uncertain or erroneous performance also will influence trust. To begin, an absence of reasons or explanations will be detrimental to trust. Dzindolet et al. [8] demonstrated that users distrust even a generally high-performing system unless provided with reasons for why performance errors have occurred. Moreover, providing these reasons can maintain or even increase trust even when the system performs poorly, as long as the explanations do not evoke counter-trust attributions. Two attributions arising from malfunctions or errors that threaten trust are incompetence and betrayal. Deception or betrayal will break trust more irrecoverably than performance errors or incompetence. Consequently, uncertainty about honesty or benign "intent" will endanger trust more than uncertainty about performance or performative competence.

The impact of errors or malfunctions also will depend on the extent to which they can be rectified or undone. Uncertainties regarding reversible or steerable decisions are less detrimental to trust than uncertainties about irrevocable decisions [35]. Smithson and Ben-Haim [40] argue that steerable or revocable choices are more robust under extreme uncertainty than irrevocable ones. One aspect of their robustness is that such choices engender less fear of unknowns and thereby pose less of a threat to trust relations.

10.2.3 Trust and Dilemmas

Trust may involve dilemmas, which arise from particular sources of uncertainty and generate additional unknowns. Here, "dilemmas" refer to situations in which multiple rational actors' pursuit of self-interests lead to sub-optimal joint outcomes. Recently

attention has been given to the "driverless car dilemma": People want others to have driverless cars programmed to sacrifice its passenger for the greater good, but they do not prefer those cars for themselves [3]. Viewed from the utilitarian assumption that sacrificing the passenger for the "greater good" is a public good regardless of whether the passenger is oneself or another person, this is a classical free-rider dilemma.

One line of reasoning about rational self-interest suggests that trust itself is inherently dilemmatic. The so-called "trust game" [2] has spawned a large literature. The original two-player procedure involves two stages. Both players are given an initial endowment of \$10, one player is assigned to be the "sender", and the other assigned to be the "receiver". In the first stage, the sender passes any amount, \$0$< s <$\$10, to the receiver. The sender retains \$10 - s, and the experimenter triples the amount sent, with 3 s passed to the receiver. In the second stage, the receiver passes any amount of the money received \$ 0 $< r <$\$3s, back to the sender. The amount passed by the sender is supposed to measure trust, and the amount returned by the receiver to measure trustworthiness. A self-interested rational sender or receiver should send nothing, and therein lies the dilemma claim. However, human players regularly demonstrate willingness to send sizeable amounts (see [21] for a meta-analysis of 162 experimental studies showing that this finding is robust across 35 countries).

Even if one does not accept the notion that trust is dilemmatic, dilemmas can pose problems for human trust in automatons that are programmed to be rational utility-maximizers. It is not difficult to imagine social dilemmas that could confront automatons and their human teammates in military combat. Suppose that enemy automatons A and B consider two alternative strategies available to each of them, A1 and A2 versus B1 and B2. To simplify matters, suppose that the stakes are the loss of 1000 lives on either or both sides. Both automatons are programmed to value the magnitude of utility for own-side casualties as 4 times greater than enemy-side casualties. That is, the utility of one own-side casualty is -4 whereas the utility of an enemy-side casualty is +1.

If A chooses A_1 and B chooses B_1 then A estimates a probability of $1/2$ of 1000 A-side casualties but also estimates 1000 B-side casualties for sure, so the expected utility for A is $U_{a11} = -4K/2 + 1K = -1K$. For the same combination of strategies, B also estimates a probability of $1/2$ of 1000 A-side casualties but only a probability of $1/2$ of 1000 B-side casualties, for an expected utility of $U_{b11} = 1K/2 - 4K/2 = -1.5K$. The remaining expected utilities are as follows.

For the $A_1 - B_2$ combination, $U_{a21} = -4K/2 = -2K$ and $U_{b21} = 1K/2 = 0.5K$;

For the $A_2 - B_1$ combination, $U_{a12} = 1K$ and $U_{b12} = -4K/2 = -2K$;

For the $A_2 - B_2$ combination, $U_{a22} = -4K + 1K/2 = -3.5K$ and $U_{b22} = 1K - 4K = -3K$.

These expected utilities are displayed in the upper half of Table 10.1 in units of 1000, with the appropriate row and column sums. The sums reveal that automaton A will conclude that A2 is its best strategy and automaton B will conclude that B2 is its best strategy. The result is the worst expected outcomes for both of them. This is a Chicken Game structure. Choosing any other combination instead would benefit both sides.

Table 10.1 A two-automaton dilemma

		A		
		A_1	A_2	
B	B_1	-1 -1.5	1 -2	-3.5
	B_2	-2 0.5	-3.5 -3	-2.5
		-3.0	-2.5	

$$
\begin{array}{ll}
A_1 : U_{a11} = -V_{aa}/2 + V_{ab} & A_2 : U_{a12} = V_{ab} \\
B_1 : U_{b11} = -V_{ba}/2 - V_{bb}/2 & B_1 : U_{b12} = -V_{bb}/2 \\
A_1 : U_{a21} = -V_{aa}/2 & A_2 : U_{a22} = -V_{aa} + V_{ab}/2 \\
B_2 : U_{b21} = V_{ba}/2 & B_2 : U_{b22} = V_{ba} - V_{bb}
\end{array}
$$

The lower half of Table 10.1 displays the utility formulas, where V_{aa} is the value given to A-side casualties by A, V_{ba} is the value of A-side casualties for B, V_{ab} is the value of B-side casualties for A, and V_{bb} the value of B-side casualties for B. Straightforward algebraic arguments show that regardless of the positive numbers assigned to these valuations, the $A_2 - B_2$ combination always is chosen by automatons A and B. Moreover, it is easy to show that this choice always is sub-optimal for both of them (even if it is not always the worst), because $U_{a21} > U_{a22}$, $U_{b12} > U_{b22}$, and $U_{b11} > U_{b22}$. Finally, it is clear that this structure always is a Chicken Game because the best outcome for A always is the $A_1 - B_2$ strategy combination whereas for B it is the $A_2 - B_1$ combination.

The prospect of such dilemmas raises a problem of trust in automatons for their human teammates and/or operators. How are they to know when, or how often, dilemmas like this will arise, and what can be done about them when they do? The obvious solutions, such as engaging in honest communication with the enemy automaton, often are not available in military situations as they may be for networked driverless cars.

10.3 Factors Affecting Human Reactivity to Risk and Uncertainty, and Trust

In this section, we survey factors affecting tolerance of uncertainties. These factors come in three kinds: the nature of the uncertainties themselves and how humans differentiate among varieties of unknowns, the psychological dispositions that influence tolerance of unknowns in general, and the conditions in groups or organizations that influence norms regarding the treatment of unknowns.

10.3.1 Kinds of Uncertainty, Risks, Standards, and Dispositions

Humans think and act as though there are distinct kinds of unknowns. They regard some kinds as worse than others, and may trade one kind for a more preferred kind. People's risk perceptions can be modulated by influences such that those perceptions will not match so-called "objective" risk assessments. They also may apply different standards of proof to different settings, and the burden of proof will depend on the assumptions they have made. Likewise, humans vary in their orientations toward and tolerance of risks and unknowns. All of these considerations are relevant to trust in HRI settings, and this section reviews them with this in mind. Starting with probabilities, there is ample evidence that human reactivity to probabilities is not linear in the probabilities, even when those probabilities are accurate. People tend to over-weight risks that have small probabilities, particularly if the stakes are high, and they have difficulty making meaningful decisional distinctions between small probabilities, even when these differ by orders of magnitude (such as one in a million versus one in ten thousand). They do, however, make a strong distinction between a probability of 0 and a very small nonzero probability. Trust in an automaton therefore is unlikely to be improved noticeably by decreasing the probability of automaton failure from, say, one in ten thousand to one in a hundred thousand. However, it is likely to increase substantially if the probability of failure is reduced from one in a hundred thousand to zero.

A relevant body of work here is on the relationship between judgments of probabilities and sample space partitions [13]. This line of research has shown that people anchor on the number of outcomes that is salient to them when making probability judgments. If they think in terms of K possible outcomes (i.e., a K-fold sample space partition), then they will anchor on probabilities of $1/K$ for each of the outcomes, and then adjust away from that when presented with relevant information. Smithson and Segale [41] demonstrated that partition-dependency effects hold even when people are using imprecise probabilities (e.g., probability intervals). An implication is that trust in an automaton can be influenced by priming users to consider its performance outcomes in alternative partitions. For instance, unpacking good outcomes into $K - 1$ sub-categories ($K > 2$) but lumping bad outcomes together into one category will anchor users on $1/K$ probability of a bad outcome, whereas packing both good and bad outcomes into one category will anchor users on a probability of 1/2 for a bad outcome.

Turning now to types of unknowns, there are long-running debates among proponents of formal frameworks for uncertainty about whether all uncertainties can be handled by some version of probability theory. These debates will not be surveyed here, but one of the motivations for them has been evidence of widespread human intuitions that not all uncertainties are probabilistic. Instead, research in judgment and decision making under uncertainty has revealed that uncertainty arising from ambiguous or conflicting information influences judgments and decisions in ways that probabilistic uncertainty does not. Ambiguity has been widely studied in psy-

chology and economics, beginning with Ellsberg's [10] seminal paper in which he demonstrated that people prefer a gamble with precisely specified probabilities to a gamble with imprecise probabilities, although the expected utilities for both gambles are identical. Although ambiguity aversion is not universally observed under all conditions (ambiguity-seeking may be observed, for example, for very low probabilities), the key point here is that people behave as though ambiguity is a different kind of uncertainty from probability that is relevant in their decisions. Several studies of uncertainty arising from conflicting information have found that there is a greater aversion to conflicting information than to ambiguous information [1, 4, 5] (e.g., [36]). Conflict aversion has been manifested in two ways. First, a majority of people prefer to receive or deal with messages from ambiguous rather than conflicting sources of information (see [36, 38]). Second, people tend to make more pessimistic estimates for future outcomes under conflict than ambiguity [4, 5, 38].

These findings suggest that ambiguous and conflicting signals or indications from an automaton may have different impacts on trust. These distinctions have implications for trust in HRI. Among the demonstrations [36] regarding conflict aversion is the finding that people usually assume that experts or computer models should agree in their forecasts and diagnoses. They prefer ambiguous but agreeing forecasts over unambiguous but disagreeing ones, even when these are informationally equivalent. Importantly, they attribute less trustworthiness to disagreeing experts or expert systems than to ambiguous but agreeing ones. It therefore seems plausible that ambiguous but agreeing signals or performance indicators from a single automaton will be less detrimental to trust than unambiguous but conflicting signals or indicators. If true, an example of a practical application is in the design of failure-mode indicators for an automaton whose operation is to be halted by a human overseer if failure is sufficiently indicated. A risk-averse approach would be to design the automaton's failure-mode indicators to be "trigger-happy" in the sense that at least one of them is likely to indicate possible failure even under a low probability that a malfunction has occurred.

The conflict versus ambiguity distinction also has implications for teams with multiple networked automatons and humans, in which the automatons are providing multiple assessments or predictions regarding the same situation. Unambiguous but disagreeing forecasts will be more detrimental to trust of the ensemble of automatons than ambiguous but agreeing ones. They also are likely to cause greater risk-aversion in the human team members. Another important kind of uncertainty is sample space ignorance, whereby the decision maker does not know all of the possible outcomes. With complex software, for instance, it is a commonplace for even its coders not to know all of its possible failure modes. Sample space ignorance has been shown in at least one study to be aversive [39]. To my awareness, no work has been done on the impact of sample space ignorance on trust. Nonetheless, it seems plausible that automatons will be viewed by users as more trustworthy if all of their possible failure modes are known than if users believe that these modes are not completely known.

What characteristics of risks besides probabilities influence human perceptions of riskiness? A large body of research on this topic indicates that people react most strongly to those risks that are hard to understand, involuntary, and invisible [22].

Typical examples are risks associated with nuclear power, nanotechnology, and climate change. Strong fears may persist despite evidence and reassurances by experts that a particular risk is minimal or unlikely. On the other hand, people are likely to be overly complacent about risks that are familiar, voluntary, and visible. Examples of this kind of risk include driving an automobile, handling or using a firearm, and using power-tools.

An additional relevant, but often neglected, characteristic of risks is whether the relevant unknowns are reducible or not. Reducible unknowns may be less corrosive of trust than irreducible ones, especially if there are measures in place to eventually eliminate these unknowns. As AI becomes more complex, irreducible uncertainties about automaton behavior will become more commonplace and may pose an obstacle to building trust in HRI.

The burden of proof identifies the party or position that must build a case to overturn a default position. (e.g., the presumption of innocence in a Western court trial places the burden of proof on the prosecution). Trust can be presumed, in cases such as role-based trust where the role involves expertise and the experts have been certified as qualified to perform the role. Given the current state of the art in HRI, presumed trust seems unlikely and so the burden of proof most often will fall on the technology and the automaton that instantiates it. However, as automatons become more advanced and more human-like, automatons may be increasingly presumed trustworthy until they prove otherwise. This prospect adds a new twist to considerations of what constitutes "appropriate" trust.

The standard of proof refers to the strength and weight of evidence required for a case to be regarded as "proven". In Western criminal trials, the conventional standard of proof is evidence of guilt "beyond reasonable doubt", whereas in civil cases the standard is "on the balance of probabilities". Standards of proof therefore demarcate thresholds for tolerance of uncertainty. Differing standards of proof regarding automaton trustworthiness between their designers and users will raise problems, so establishing agreements about such standards will be an important aspect of automaton development, testing, and deployment.

Finally, psychological dispositions may play a role in building trust. Some people are less trusting than others, they may be more risk-averse, and/or more intolerant of uncertainty. Dispositions such as these may influence the standard of proof a human brings to HRI when making judgments of automaton trustworthiness. Only few HRI studies have systematically investigated the role of human-related characteristics (e.g. level of expertise, personality traits such as extroversion [17]) and environmental factors (e.g. culture, task type [27]). To my knowledge, none have investigated the role of trait-level trustingness, risk orientation, or tolerance of uncertainty regarding their influences on the nature of trust in HRI. Because trust relations are strongly context-dependent, it is possible that psychological traits will not have a strong influence here, but this possibility has yet to be ascertained.

10.3.2 Presumptive and Organizational-Level Trust

Kramer and Lewicki [24] introduce the notion of "presumptive" trust as a kind of depersonalized basis for trust that has more to do with indirect indicators such as reputation and properties of organizational or group settings such as shared identity, common fate, and interdependence, than with direct indicators of trustworthiness as manifested by the potential trustee. The term "presumptive" conveys that this kind of trust is a default stance on the part of the trustor, and often operates in a tacit way. According to Kramer and Lewicki, presumptive trust has at least one of three primary bases: Identities, roles, and rules.

Identity-based trust is the expectation that fellow in-group members can be trusted, and some scholars have argued that this is based on an expectation of general reciprocity within the boundaries of the in-group [12]. Shared identity is unlikely to be a basis for human trust of automatons, although it certainly is plausible that "in-group" automatons may be trusted more than "out-group" automatons, even when both categories of automaton are "on the same side".

Role-based trust probably would better be thought of as "system-based". The primary idea here is that an individual occupying a specific role in an organization may be trusted because both the nature of the role and the system of training and/or selecting people to occupy that role are trusted. Thus, we will trust a robot if we trust robotics and also trust the engineering programs that train roboticists. Or, we may trust a particular brand of automaton because we trust that particular company and its selection processes for hiring engineers and programmers.

Rule-based trust has its source in the codified norms and other rules for behavior within a group or organization, and the expectation that members have been socialized to follow the rules and adhere to the norms. "Honour" codes are an example of this kind of trust basis. Analogs for this kind of trust in HRI include beliefs about the robot's adherence to its programmed protocols, and compatibility between those protocols and human social and psychological norms. There may be a design tradeoff here between a preference for robots that "blindly" adhere to their inbuilt protocols and a preference for robots whose behavior is flexible and adapts to novel situations.

Risk management norms in a group or organization will influence the development of trust in HRI. Perhaps the most obvious kind of influence stems from the "tightness" of the organizational culture [14]. So-called "tight" cultures have numerous strong norms and very little tolerance of deviant behavior, whereas "loose" cultures' social norms are relatively weak and they are permissive of deviant behavior. Research into this cultural dimension has found a correlation between tightness and the magnitude of risks in the ecology occupied by a culture. This connection suggests that tighter cultures will be more risk-averse and less trusting. While the research program elaborated by [14] has focused on national cultures, it is plausible that these same connections and the tightness construct will apply to organizations and groups.

10.3.3 Trust Repair

Kramer and Lewicki [24] observe that most approaches to trust repair have only focused on changing cognitions, thereby neglecting emotional or behavioral aspects of trust repair. Much of this research also has emphasized routes to repair that may not apply in HRI, although as automatons are increasingly humanized more of these routes may become available. Also, it is arguably an open question as to whether some apparently incongruous acts by an automaton could nevertheless aid in trust repair. For example, would an apology by a robot for its error assuage human users?

Both explanations and apologies have been found to help restore trust, but generally if accompanied by some actual reparations or measures to prevent further breaches of trust. Tomlinson, et al. [42] investigated the characteristics of apologies influencing their effectiveness in trust repair. They found that an apology was more effective if issued sooner than later after a breach of trust. They also found that apologies and explanations that had the trust violator taking responsibility for the breach were more effective than accounts that blamed other parties or external factors for the breach. A possible exception to this finding, pointed out by [24], is when the breach involves a violation of integrity. In that case, being able to deny responsibility for such a violation may be more effective.

Penance and reparations have been extensively studied in regard to trust repair. One problem for HRI is that, like apologies, penance and reparation on the part of an automaton may be largely irrelevant unless humans have anthropomorphized the automaton to the extent that they attribute emotional responses to it. However, such measures could be applied to the designers or producers of the automaton, especially if trust in the automaton is primarily a matter of trust in its designers and/or producers.

Similar arguments apply to other more "legalistic" trust repair mechanisms, such as rules, contracts, monitoring systems, and sanctions against further trust violations. Most of these are attempts to ensure that the trusted party is motivated not to breach trust again, which is irrelevant to an automaton unless its users attribute motivations to it. One partial exception to this is reinforcement schedules in machine learning, which could be revised in the service of preventing further malfunctions or errors by the automaton.

10.4 Concluding Remarks

In this chapter, we have surveyed the following factors in HRI that influence the nature and development of trust:

- The scope of an automaton's capabilities and responsibilities, and the extent of its control over decisions
- Whether the sources of potential malfunctions or mishaps are internal or external to the automaton
- Which kinds of errors or malfunctions are most important or consequential

- The impact of uncertainty about benign intent versus competence or reliability
- Uncertainties arising from the prospect of social dilemmas involving interacting automatons, especially opponent automatons
- Organization-based trust and the impact of organizational norms and culture
- Factors influencing trust repair when trust has been eroded or lost.

This chapter also has provided suggestions for several avenues of further research and theoretical developments regarding the role of uncertainties in HRI, specifically in connection with trust. A major theme of this chapter is that almost all treatments of uncertainty in relation to matters of trust have over-simplified both the role and nature of uncertainty. Regarding its role, on the one hand, it is widely claimed that trust serves to reduce uncertainty. On the other, it also is widely claimed that uncertainty is endemic in a trust relationship. Absent from these accounts is the realization that in establishing a trust relationship, the trustor is trading the reduction of one set of uncertainties for the creation of another set of uncertainties. Typically, the tradeoff involves reducing uncertainty about outcomes (to be attained by the trustee) at the expense of tolerating uncertainty about the means by which the trustee pursues and achieves those outcomes. Likewise, the role of uncertainty in distrust has not been fully understood, especially in regard to the license for secrecy, deception, and other forms of ignorance production that distrust provides for the distrustor.

Uncertainty also has largely been treated as if it is unitary or monolithic, and a "negative" that people are motivated to be rid of. These over-simplifications persist throughout both the human sciences and engineering. People have uses for unknowns and unknowns underpin important forms of social capital, as is exemplified by the fact that a trust relationship is predicated on tolerated ignorance. Likewise, as has been clearly articulated in this chapter, people think and act as though there are different kinds of uncertainty, and as though those differences are important. For instance, they prefer agreeing but vague experts to precise but disagreeing experts (i.e., "conflict aversion"), and they trust the former more than the latter. The impacts of different kinds of uncertainty on trust in HRI remain to be systematically investigated, but this chapter points to clear directions for such research.

References

1. A. Baillon, L. Cabantous, P.P. Wakker, Aggregating imprecise or conflicting beliefs: an experimental investigation using modern ambiguity theories. J. Risk Uncertainty **44**(2), 115–147 (2012)
2. J. Berg, J. Dickhaut, K. McCabe, Trust, reciprocity, and social history. Game Econ. Behav. **10**(1), 122–142 (1995)
3. J.-F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles. Science **352**(6293), 1573–1576 (2016)
4. L. Cabantous, Ambiguity aversion in the field of insurance: insurers' attitude to imprecise and conflicting probability estimates. Theor. Decis. **62**(3), 219–240 (2007)

5. L. Cabantous, D. Hilton, H. Kunreuther, E. Michel-Kerjan, Is imprecise knowledge better than conflicting expertise? evidence from insurers' decisions in the united states. J. Risk Uncertainty **42**(3), 211–232 (2011)
6. D.C. Dennett, *The Intentional Stance* (MIT press, Cambridge, 1989)
7. J.J. Dijkstra, W.B.G. Liebrand, E. Timminga, Persuasiveness of expert systems. Behav. Inf. Technol. **17**(3), 155–163 (1998)
8. M.T. Dzindolet, S.A. Peterson, R.A. Pomranky, L.G. Pierce, H.P. Beck, The role of trust in automation reliance. Int. J. Hum.-Comput. Stud. **58**(6), 697–718 (2003)
9. M.T. Dzindolet, L.G. Pierce, H.P. Beck, L.A. Dawe, B.W. Anderson, Predicting misuse and disuse of combat identification systems. Mil. Psychol. **13**(3), 147 (2001)
10. D. Ellsberg, Risk, ambiguity, and the savage axioms, in *The Quarterly Journal of Economics*, pp. 643–669 (1961)
11. G.A. Fine, L. Holyfield, Secrecy, trust, and dangerous leisure: generating group cohesion in voluntary organizations. Soc. Psychol. Q. pp. 22–38 (1996)
12. M. Foddy, T. Yamagishi, *Whom Can We Trust*, chapter Group-based trust (Russell Sage Foundation, New York, 2009), pp. 17–41
13. C.R. Fox, Y. Rottenstreich, Partition priming in judgment under uncertainty. Psychol. Sci. **14**(3), 195–200 (2003)
14. M.J. Gelfand, J.L. Raver, L. Nishii, L.M. Leslie, J. Lun, B.C. Lim, L. Duan, A. Almaliach, S. Ang, J. Arnadottir et al., Differences between tight and loose cultures: a 33-nation study. Science **332**(6033), 1100–1104 (2011)
15. P.A. Hancock, D.R. Billings, K.E. Schaefer, Can you trust your robot? Ergon. Des. **19**(3), 24–29 (2011)
16. R. Hardin, The street-level epistemology of trust. Analyse & Kritik **14**(2), 152–176 (1992)
17. K.S. Haring, Y. Matsumoto, K. Watanabe, How do people perceive and trust a lifelike robot, in *Proceedings of the World Congress on Engineering and Computer Science*, Vol. 1 (2013)
18. L.T. Hosmer, Trust: the connecting link between organizational theory and philosophical ethics. Acad. Manag. Rev. **20**(2), 379–403 (1995)
19. B. Huebner, Commonsense concepts of phenomenal consciousness: does anyone care about functional zombies? Phenomenol. Cogn. Sci. **9**(1), 133–155 (2010)
20. J.-Y. Jian, A.M. Bisantz, C.G. Drury, Foundations for an empirically determined scale of trust in automated systems. Int. J. Cogn. Ergon. **4**(1), 53–71 (2000)
21. N.D. Johnson, A.A. Mislin, Trust games: a meta-analysis. J. Econ. Psychol. **32**(5), 865–889 (2011)
22. R.E. Kasperson, O. Renn, P. Slovic, H.S. Brown, J. Emel, R. Goble, J.X. Kasperson, S. Ratick, The social amplification of risk: a conceptual framework. Risk Anal. **8**(2), 177–187 (1988)
23. R.M. Kramer, Trust and distrust in organizations: emerging perspectives, enduring questions. Annu. Rev. Psychol. **50**(1), 569–598 (1999)
24. R.M. Kramer, R.J. Lewicki, Repairing and enhancing trust: approaches to reducing organizational trust deficits. Acad. Manag. Ann. **4**(1), 245–277 (2010)
25. J. Langan-Fox, M.J. Sankey, J.M. Canty, Human factors measurement for future air traffic control systems. Hum. Factors: J. Hum. Factors Ergon. Soc. **51**(5), 595–637 (2009)
26. J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance. Hum. Factors: J. Hum. Factors Ergon. Soc. **46**(1), 50–80 (2004)
27. D. Li, P.L. Patrick Rau, Y. Li, A cross-cultural study: effect of robot appearance and task. Int. J. Soc. Robot. **2**(2), 175–186 (2010)
28. P. Madhavan, D.A. Wiegmann, Similarities and differences between human-human and human-automation trust: an integrative review. Theor. Issues Ergon. Sci. **8**(4), 277–301 (2007)
29. D.J. McAllister, Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. Acad. Manag. J. **38**(1), 24–59 (1995)
30. D.H. McKnight, N.L. Chervany, Trust and distrust definitions: one bite at a time, in *Trust in Cyber-societies* (Springer, Berlin, 2001), pp. 27–54

31. K.E. Oleson, D.R. Billings, V. Kocsis, J.Y.C. Chen, P.A. Hancock, Antecedents of trust in human-robot collaborations, in *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (IEEE, 2011), pp. 175–178
32. S. Ososky, D. Schuster, E. Phillips, F.G. Jentsch, Building appropriate trust in human-robot teams, in *2013 AAAI Spring Symposium Series* (2013)
33. J.B. Rotter, Generalized expectancies for interpersonal trust. Am. psychol. **26**(5), 443 (1971)
34. J.B. Rotter, Interpersonal trust, trustworthiness, and gullibility. Am. psychol. **35**(1), 1 (1980)
35. M. Salem, G. Lakatos, F. Amirabdollahian, K. Dautenhahn, Would you trust a (faulty) robot?: effects of error, task type and personality on human-robot cooperation and trust, in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, New York, 2015) pp. 141–148
36. M. Smithson, Conflict aversion: preference for ambiguity vs conflict in sources and evidence. Organ. Behav. Hum. Decis. Process. **79**(3), 179–198 (1999)
37. M. Smithson, *Uncertainty and risk: Multidisciplinary perspectives*, chapter The many faces and masks of uncertainty (Earthscan, London, 2008), pp. 13–25
38. M. Smithson, Conflict and ambiguity: preliminary models and empirical tests, in *Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, pp. 303–310 (2013)
39. M. Smithson, T. Bartos, K. Takemura, Human judgment under sample space ignorance. Risk Decis. Policy **5**(02), 135–150 (2000)
40. M. Smithson, Y. Ben-Haim, Reasoned decision making without math? adaptability and robustness in response to surprise. Risk Anal. **35**(10), 1911–1918 (2015)
41. M. Smithson, C. Segale, Partition priming in judgments of imprecise probabilities. J. Stat. Theor. Pract. **3**(1), 169–181 (2009)
42. E.C. Tomlinson, B.R. Dineen, R.J. Lewicki, The road to reconciliation: antecedents of victim willingness to reconcile following a broken promise. J. Manag. **30**(2), 165–187 (2004)
43. D.A. Wiegmann, A. Rich, H. Zhang, Automated diagnostic aids: the effects of aid reliability on users' trust and reliance. Theor. Issues Ergon. Sci. **2**(4), 352–367 (2001)
44. R.E. Yagoda, D.J. Gillan, You want me to trust a robot? the development of a human-robot interaction trust scale. Int. J. Soc. Robot. **4**(3), 235–248 (2012)