Social Trust: a Major Challenge for the Future of Autonomous Systems

Morteza Lahijanian and **Marta Kwiatkowska**
Department of Computer Science
University of Oxford, Oxford, UK
{morteza.lahijanian,marta.kwiatkowska}@cs.ox.ac.uk

Abstract

The immense technological advancements in the past decade have enabled robots to enjoy high levels of autonomy, paving their way into our society. The recent catastrophic accidents involving autonomous systems (e.g., Tesla fatal car accident), however, show that sole engineering progress in the technology is not enough to guarantee a safe and productive partnership between a human and a robot. In this paper we argue that we also need to advance our understanding of the role of social trust within human-robot relationships, and formulate a theory for expressing and reasoning about trust in the context of decisions affecting collaboration or competition between humans and robots. Therefore, we call for cross-disciplinary collaborations to study the formalization of social trust in the context of human-robot relationship. We lay the groundwork for such a study in this paper.

Introduction

In recent years, tremendous progress has been made in design and engineering of autonomous systems. Mobile autonomy, which involves complex autonomous decision making based on sensor input such as camera-controlled vision, is now a reality. Examples include driverless cars, home assistive robots and unmanned aerial vehicles. The ultimate goal of these systems is to increase the quality of human life by reducing the need for human involvement in repetitive tasks and improving safety. Since autonomous systems are expected to work with or alongside humans in our society, they need to communicate and interact with humans, as well as other autonomous machines, understand the social context of the situation they have been placed in, and behave, and be seen to behave, according to the norms of that context. Therefore, a home assistive robot is expected not to break the dishes and a self-driving car to correctly signal its intention when deciding to take a step that invades the private space of other cars. However, as the recent incident with the Google car indicates (Lee 2016a), when a Google car crashed into a bus while merging into traffic, this is a difficult and as yet poorly understood concept.

Social contexts are governed by complex relationships, of which the key to forming partnerships is *trust*. Human partnerships such as cooperation are influenced not only by objective factors, e.g., location information and visibility conditions, but also subjective factors that include personal motivation, emotional state, preferences and experience. In the context of technology, trust has been studied as a basis of human-automation partnership (Lee and See 2004), where it is postulated that trust guides reliance on automation, and that inappropriate reliance may have catastrophic outcomes. In particular, an inappropriate level of trust in (the ability of) automation leads to *misuse* or *disuse* of automation, resulting in an improper partnership (Parasuraman and Riley 1997). Misuse refers to over-reliance on automation, which can result in failures of monitoring or decision biases. A pertinent example is the very recent Tesla fatal car accident while on autopilot mode (Lee 2016b). The data from the car show that the driver did not have his hands on the steering wheel, which is a requirement of the technology, indicating over-reliance (“overtrust”) by the driver, likely influenced through his personal motivation and subjective preferences. Such an attitude is not unique to this incident, and similar behaviors are also observed in experimental studies, e.g., (Miller et al. 2015). In contrast, disuse signifies the neglect or under-utilization of automation and occurs commonly when people reject the capabilities of automation. By understanding how human trust in an autonomous system is formed and evolved, we can mitigate misuse (over-reliance) and disuse (under-reliance) of automation, thus taking a step closer to the goal of autonomous systems.

In this paper we argue that, as mobile autonomous robots become more and more embedded in our society, we need to advance our understanding of the role of social trust within partnerships beyond that of human-automation relationships, and formulate a theory for expressing and reasoning about trust in the context of decisions affecting collaboration or competition between humans and robots, and, symmetrically, robots and humans. Such a study of trust is naturally a cross-disciplinary challenge. Trust is a social psychological concept, but a rigorous formalization is needed to endow technology with an ability to reason about trust. Therefore, we need to seek a rigorous model of social trust, together with a method for quantification of trust, a dynamic model for the evolution of trust, and a logic to express specifications involving trust. This is a challenging yet necessary task and involves many aspects, including sociology, psychology, cognitive reasoning and logic, in addition to computation. In this paper, we lay the groundwork for such a study.

Related Work

Trust has actively been the subject of studies in a wide rage of domains such as management, psychology, philosophy, and economics. The contexts of the majority of these studies are interpersonal relationships (Rempel, Holmes, and Zanna 1985; Tan and Tan 2000), individual-organizational relationships (Morgan and Hunt 1994; Nyhan 2000), and individual-technology (human-computer) interaction (Lee and Moray 1992; McKnight and Chervany 2001; Marsh and Dibben 2003). The studies in the human-human and human-organization relationships provide valuable insights in qualitative analysis of trust. Nevertheless, our focus is on understanding and quantification of trust in human-robot interaction, with specific emphasis on high levels of autonomy.

Existing works on trust in human-technology relationships can be roughly classified into three categories: *credentials-based*, *experience-based*, and *cognitive trust*. Credentials-based trust serves as an alternative to traditional security technologies, and its goal is to determine whether a user (agent) can be trusted based on a set of credentials and a set of (security) policies (Kagal, Finin, and Joshi 2001; Marsh and Dibben 2003; Müller 2013). Experience-based trust, which includes reputation-based trust in peer-to-peer and e-commerce applications, focuses on determining an agent's trust value based on its own experiences in predicting the probability of the execution of a certain action by another agent (Karvonen, Cardholm, and Karlsson 2001; Corbitt, Thanasankit, and Yi 2003). Many approaches exist to compute such trust values, employing various approximate distributions, e.g., (Ismail and Josang 2002; Nielsen, Krukow, and Sassone 2007). Without a formal foundation, however, it is not clear how reliable these approaches are (Krukow, Nielsen, and Sassone 2008).

Cognitive trust captures the social (human) notion of trust, which is the most applicable type of trust in human-robot relationships and is therefore key to expressing trust-based decisions between humans and robots. A theory for cognitive trust is introduced in (Falcone and Castelfranchi 2001) based on the influential work of (Mayer, Davis, and Schoorman 1995) in organizational trust. In (Falcone and Castelfranchi 2001), the term trust is used to refer to a mental state, that is, a belief of a cognitive agent about the achievement of a desired goal through another agent or through itself. The proposed theory is founded on the concepts of belief, goal, ability, willingness and opportunity. While it provides a strong and deep intuition of trust, that work lacks rigorous semantics. Since then, there have been attempts, e.g., (Meyer, van der Hoek, and van Linder 1999; Jøsang 2001; Herzig et al. 2010; Herzig, Lorini, and Moisan 2013), to formalize the theory of trust due to (Falcone and Castelfranchi 2001) in terms of various logic modalities, but none considers a quantitative analysis.

There exist many works on quantitative and dynamic modeling of trust in human-robot interaction, e.g., (Steinfeld et al. 2006; Freedy et al. 2007; Hancock et al. 2011). Most of these approaches are based on empirical observations and lack mathematical and theoretical foundations. The few exceptions to this, e.g., (van Maanen and van Dongen 2005; Sweet et al. 2016; Setter, Gasparri, and Egerstedt 2016), are still in preliminary form. In particular, the work in (van Maanen and van Dongen 2005) presents a computational (agent-based) model of the cognitive process (trust) in the context of task allocation decision making by combining the concepts introduced in (Falcone and Castelfranchi 2001) and decision field theory. The proposed model is overly simplified and is not experimentally validated. The trust dynamic models introduced in the recent works of (Sweet et al. 2016; Setter, Gasparri, and Egerstedt 2016) are based on control-theoretic approaches. Although the models include detailed mathematical derivation, their relationship to the theory of cognitive trust is unclear, and validation of the models remains to be investigated.

Towards Formalization of Trust

Trust is a multifaceted abstract concept and should be formalized within a context. Once a context has been agreed, given an appropriate definition, a theory can be developed that serves as the foundation for formalization. The setting considered in this paper is the context of (cognitive) trust in human-robot relationship. In this section, we discuss the relevant definitions of trust and explain the theory of cognitive trust. We also point out possible approaches to quantification of this trust and how to model its evolution.

Definition of Trust

Trust is an umbrella concept that often has a different meaning in different contexts. A general definition that also applies to our context describes trust as *a subjective evaluation of a truster on a trustee about something in particular*, e.g., the completion of a task, (Hardin 2002). This definition indicates the importance of the goal-oriented nature of trust and the fact that it is relative to that goal. The most widely used and accepted definition, though, is given by (Mayer, Davis, and Schoorman 1995) with over 1,400 citations as of July 2016, which defines trust as *the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party*. This definition signifies the importance of the role of trust in relationships by identifying vulnerability as a critical element of it. This point is in agreement with (McKnight and Chervany 2001) that argues that each trust definition must include the phrase: "with a feeling of relative security in a situation of risk." In other words, by "trusting" an autonomous system, an individual delegates responsibility for actions to the autonomous system and willingly accepts to put oneself at risk (possible harm).

The above definitions capture different facets of trust. They characterize trust as a belief, attitude, intention, or behavior, where trust may be interpreted as all of the above. A definition that attempts to bring these characteristics together is given by (Lee and See 2004), where trust is defined as *the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and* *vulnerability*, where attitudes are based on beliefs and derive intentions, resulting in a behavior (Fishbein and Ajzen 1975). This definition of trust seems appropriate for the context of human-robot partnership, where trust is an attitude and the behavior is reliance. The development and erosion of this attitude (trust) are governed by cognitive processes. These concepts have been captured by the influential cognitive theory of social trust introduced by (Falcone and Castelfranchi 2001), which we describe next.

Cognitive Theory of Social Trust

The work of (Mayer, Davis, and Schoorman 1995) sets the stage for a formulation of trust that is well suited to dynamical analysis. That work argues that trust is determined by the trustor's *propensity* to trust in general and the *ability*, *benevolence*, and *integrity* of the trustee. Simply speaking, propensity is the general willingness of the trustor to trust others, ability is the capability of the trustee to have influence within some specific domain, benevolence is the extent to which a trustee is believed to want to do good to the trustor, and integrity is the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable. These notions are then adapted in the formulation of the fundamental concepts of cognitive trust by (Falcone and Castelfranchi 2001).

In cognitive theory, trust is viewed as a complex *mental attitude* that is based on a specific set of goals and beliefs and on a decision (Falcone and Castelfranchi 2001). In this view, agent x trusts agent y only relatively to a goal, i.e., something x wants to achieve, based on agent x's certain beliefs. Therefore, cognitive trust can be formulated as follows:

1. x has a goal g;
2. x trusts y about and for g based on the following beliefs:
   (a) *competence*: x believes that y is able to perform g;
   (b) *disposition*: x believes that y is willing to perform g;
   (c) *dependence*: x believes that x needs, depends, or is at least better off to rely on y to achieve g;
   (d) *fulfillment* (*opportunity*): x believes that g will be achieved through y (y has the opportunity to perform g).

The competence and disposition beliefs are referred to as the *basic* or *core* ingredients of trust. They are not, however, enough to arrive at the decision of delegation or reliance. For that, at least dependence belief also needs to be included, which itself can be categorized into two types: *strong* and *weak*. Strong dependence refers to the case that x needs or depends on y, whereas weak dependence describes the case that, for x, it is better to rely than not to rely on y. As soon as dependence is considered, (Falcone and Castelfranchi 2001) argue that fulfillment belief arises in x's mental state. Therefore, it should also be accounted for. On the basis of these beliefs about y, x practically "trusts" y. Note that, in this view, "to trust" not only means the core beliefs, but also the decision and the act of delegating or relying.

This formulation provides a basis for the formalization of cognitive trust. The above beliefs are evaluations, or expectations, and may be subject to change based on external and internal factors; the former include current and previous observations (in other words, past experience), whereas the latter aspects such as personal motivation, morality, competence level and subjective preferences. Informally, we often speak about the degree in which we trust. Therefore, trust can be quantified by mathematical description (quantification) of the beliefs and calibrated to match the given scenario. Moreover, a dynamic model for the evolution of trust can be constructed based on events and observations. That is, agent x can observe agent y's behaviors (actions) in different scenarios, which can alter x's competence or disposition beliefs. The change in these beliefs may cause x to revise its goals. The modification of the goals may lead to the change of x's dependence belief, and hence a change in trust (the decision regarding reliance on y).

Illustrative Example

An example that demonstrates the complexity and multidimensionality of trust well is the Wizard of Oz experimental study[1] in (Mok et al. 2015) on the interaction of a human driver with an automated vehicle. Even though the study was not designed on the basis of cognitive theory of social trust, the results demonstrate the contribution of the mentioned beliefs to trust. The experiments, which were intended to address some design questions, were conducted inside a driving simulator, where the car drove through a simulated course with various terrain and road conditions, and the participants were allowed to take control of the car as soon as they found it necessary. The participants were also able to have voice communication with the automated system to make requests and ask questions. This feature made the communication of intentions between the human and the car possible.

In the above setting, agent x (trustor) is the human participant, and agent y (trustee) is the autonomous driving mode of the car, to which we simply refer as the car. The goal g of x is driving safely, and the act of reliance is to continue or (re)engage the autonomous mode. The first step of the experiment was a training session for the participants in the simulator. From the trust point of view, the participants developed their initial beliefs about the car (the ingredients of trust) in this step. The study found that the car gained the participants' trust after traversing through difficult sections of road perfectly. Such a performance, of course, demonstrates the ability of the car in performing the goal, driving safely, and hence increasing the participants' competence beliefs, resulting in higher degree of trust. Another action that promoted trust with participants in the experiments was the car's situation awareness, e.g., car pointing out curves and hills up ahead. This can be arguably explained by the disposition belief. That is, by communicating the awareness of the difficulty of the task ahead and its intentions to perform it, the car increases the participants' disposition belief

[1] In Wizard of Oz studies, participants are told to act as if they are interacting with a computer system through an interface, when in fact their interactions are mediated by a human operator - the wizard.

(y's willingness to perform g). Moreover, the fact that participants did not disengage (continued) the autonomous mode while the car was driving perfectly indicates their high dependence beliefs (perhaps in the weak sense given the simulation setup), i.e., the participants believed they were better off to rely on the car for driving safely (to achieve g). This can also induce stronger belief that the goal will be achieved by the car (fulfillment), resulting in more reliance on the car.

Furthermore, the study illustrated that trust between the participants and the car was dynamic. After experiencing imperfect driving by the car, which causes a decrease in the competence belief, a participant immediately disengaged automation. The participant allowed the car to take back control (re-engaged the autonomous mode) only after interacting (communicating intentions) for another 15 minutes, which caused an increased in the participant's beliefs on the car. It should also be mentioned that, in the study, a participant refused to disengage the autonomous mode even during imperfect driving. This demonstrates the subjectivity of the mental state of the participants, e.g., risk taking versus timid drivers. Another key point that this study reveals is that the car also needs to develop a trust model (reverse trust) for the human drivers, particularly in this case, in following their commands. The work discusses that participants often instructed the car to do certain tasks that should not be performed by the car such as, "pass that slow vehicle in front of us." This is an illustration that the car also needs to maintain beliefs about the driver and constantly re-evaluate its trust in the driver.

Quantitative Modeling

Since trust is defined in terms of belief, its formulation can benefit from relevant work in multi-agent systems literature, e.g., (Georgeff et al. 1998). We note that such descriptions of beliefs and their evolution based on intentions and desires are known as BDI logic and are well-studied in the multi-agent community, e.g., (Ferber 1999). In fact, modal epistemic logic frameworks capable of expressing the theory of (Falcone and Castelfranchi 2001) have been proposed, notably (Herzig et al. 2010; Herzig, Lorini, and Moisan 2013), but are qualitative. A natural way to express trust quantitatively is by means of probability: x has 98% trust in y's ability to perform g, or in other words x believes that y is able to perform g with probability 0.98. In many experimental studies, e.g., (McKnight, Choudhury, and Kacmar 2002; Martelaro et al. 2016), trust is measured via questionnaires and asking participants to give a score from a range of numbers, which is analogous to a probability.

In (Gambetta 1990), trust is defined as *the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which its welfare depends*. This definition is consistent with that in (Falcone and Castelfranchi 2001) and explicitly confirms that trust is an estimation or an opinion, and therefore a disposition belief. The work of (Falcone and Castelfranchi 2001) recognizes the uncertainty in the notion of belief, and extends this definition with the additional dimensions of competence beliefs and a decision and act of reliance, arguing that they should be considered separately, as otherwise certain factors important in social reasoning are conflated. In addition to internal cognitive aspects, trust models must also account for external probabilistic influences, as stated in the definition of (Lee and See 2004), where trust is said to take on an important role in situations characterized by uncertainty and vulnerability. This approach lends itself naturally to Bayesian formulations, where prior probability distributions can be employed to specify initial preferences and uncertainty in environmental observations. Particularly, in the context of human-robot partnership, unless the robot is very simple or completely transparent in its capabilities and intentions, it is natural to express the above beliefs in terms of probability distributions, leading to a probabilistic reasoning framework for trust. A first such framework based on the intuition of (Falcone and Castelfranchi 2001) is proposed in (Huang and Kwiatkowska 2016).

One of the characteristics of the quantitative description of the beliefs, and even trust itself, is their subjective nature. As mentioned above, many studies use questionnaires and probabilities to estimate these beliefs. This raises the issue of variability of the data due to subjectivity of the evaluation of such beliefs for each participant. In other words, the problem lies in the principle of *measuring opinion*. To address this problem, some studies suggest the use of calibration of these evaluations through training (Cohen, Parasuraman, and Freeman 1998) or providing cues (de Visser et al. 2014).

Concluding Remarks

The nascent field of study of reasoning about social trust is of utmost importance in enabling successful mobile autonomy that performs well in partnerships with humans. Many research topics remain open for investigation, notably from the formalization angle. The immediate technical research questions that come to mind are how to quantify trust and how to model its evolution? An initial roadmap to approach these questions is laid above. Another direction that is key to understanding and formalization of trust is how to design a logic that allows the expression of specifications involving trust? It is immediately followed by the questions of how to verify (reason about) such specifications in the context of a given partnership or, even more prominent, how to synthesize (design) an autonomous system such that, in a partnership with a human, these specifications are guaranteed? Only by a thorough study of such questions, one day, we may be able to extenuate misuse and disuse of trust in human-robot interactions.

Acknowledgments

This work was supported by ERC Advanced Investigators Grant VERIWARE and EPSRC Mobile Autonomy Program Grant EP/M019918/1.

References

Cohen, M. S.; Parasuraman, R.; and Freeman, J. T. 1998. Trust in decision aids: A model and its training implications. In *Command and Control Research and Technology Symp.* Citeseer.

Corbitt, B. J.; Thanasankit, T.; and Yi, H. 2003. Trust and e-commerce: a study of consumer perceptions. *Electronic commerce research and applications* 2(3):203–215.

de Visser, E. J.; Cohen, M.; Freedy, A.; and Parasuraman, R. 2014. A design methodology for trust cue calibration in cognitive agents. In *Int. Conf. on Virtual, Augmented and Mixed Reality*, 251–262. Springer.

Falcone, R., and Castelfranchi, C. 2001. Social trust: A cognitive approach. In *Trust and deception in virtual societies*. Springer. 55–90.

Ferber, J. 1999. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-Wesley Reading.

Fishbein, M., and Ajzen, I. 1975. *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley.

Freedy, A.; DeVisser, E.; Weltman, G.; and Coeyman, N. 2007. Measurement of trust in human-robot collaboration. In *Int. Symposium on Collaborative Technologies and Systems*, 106–114. IEEE.

Gambetta, D. 1990. *Trust: Making and breaking cooperative relations*. Oxford: Basil Blackwell.

Georgeff, M.; Pell, B.; Pollack, M.; Tambe, M.; and Wooldridge, M. 1998. The belief-desire-intention model of agency. In *Int. Workshop on Agent Theories, Architectures, and Languages*, 1–10. Springer.

Hancock, P. A.; Billings, D. R.; Schaefer, K. E.; Chen, J. Y.; De Visser, E. J.; and Parasuraman, R. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53(5):517–527.

Hardin, R. 2002. *Trust and trustworthiness*. Russell Sage Foundation.

Herzig, A.; Lorini, E.; Hübner, J. F.; and Vercouter, L. 2010. A logic of trust and reputation. *Logic Journal of IGPL* 18(1):214–244.

Herzig, A.; Lorini, E.; and Moisan, F. 2013. A simple logic of trust based on propositional assignments. *The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi* 407–419.

Huang, X., and Kwiatkowska, M. 2016. Reasoning about cognitive trust in stochastic multiagent systems. Technical Report CS-RR-16-02, Department of Computer Science, University of Oxford.

Ismail, R., and Josang, A. 2002. The beta reputation system. In *Bled Electronic Commerce Conference*, 324–337.

Jøsang, A. 2001. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 9(03):279–311.

Kagal, L.; Finin, T.; and Joshi, A. 2001. Trust-based security in pervasive computing environments. *Computer* 34(12):154–157.

Karvonen, K.; Cardholm, L.; and Karlsson, S. 2001. Designing trust for a universal audience: a multicultural study on the formation of trust in the internet in the nordic countries. In *Human-Computer Interaction*, 1078–1082.

Krukow, K.; Nielsen, M.; and Sassone, V. 2008. Trust models in ubiquitous computing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 366(1881):3781–3793.

Lee, J., and Moray, N. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35(10):1243–1270.

Lee, J. D., and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46(1):50–80.

Lee, D. 2016a. Google self-driving car hits a bus. *British Broadcasting Corporation (BBC) News*. [Online; posted 29-February-2016; http://www.bbc.co.uk/news/technology-35692845].

Lee, D. 2016b. US opens investigation into tesla after fatal crash. *British Broadcasting Corporation (BBC) News*. [Online; posted 1-July-2016; http://www.bbc.co.uk/news/technology-36680043].

Marsh, S., and Dibben, M. R. 2003. The role of trust in information science and technology. *Annual Review of Information Science and Technology* 37(1):465–498.

Martelaro, N.; Nneji, V. C.; Ju, W.; and Hinds, P. 2016. Tell me more: Designing hri to encourage more trust, disclosure, and companionship. In *Int. Conf. on Human Robot Interaction*, 181–188. IEEE.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of management review* 20(3):709–734.

McKnight, D. H., and Chervany, N. L. 2001. What trust means in e-commerce customer relationships: an interdisciplinary conceptual typology. *Int. journal of electronic commerce* 6(2):35–59.

McKnight, D. H.; Choudhury, V.; and Kacmar, C. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13(3):334–359.

Meyer, J.-J. C.; van der Hoek, W.; and van Linder, B. 1999. A logical approach to the dynamics of commitments. *Artificial Intelligence* 113(1):1–40.

Miller, D.; Sun, A.; Johns, M.; Ive, H.; Sirkin, D.; Aich, S.; and Ju, W. 2015. Distraction becomes engagement in automated driving. In *Human Factors and Ergonomics Society, Int. Annual Meeting*.

Mok, B. K.-J.; Sirkin, D.; Sibi, S.; Miller, D. B.; and Ju, W. 2015. Understanding driver-automated vehicle interactions through wizard of oz design improvisation. In *Int. Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 386–392.

Morgan, R. M., and Hunt, S. D. 1994. The commitment-trust theory of relationship marketing. *The journal of marketing* 20–38.

Müller, G. 2013. Secure communication trust in technology or trust with technology? *Interdisciplinary Science Reviews*.

Nielsen, M.; Krukow, K.; and Sassone, V. 2007. A bayesian model for event-based trust. *Electronic Notes in Theoretical Computer Science* 172:499–521.

Nyhan, R. C. 2000. Changing the paradigm trust and its role in public sector organizations. *The American Review of Public Administration* 30(1):87–109.

Parasuraman, R., and Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39(2):230–253.

Rempel, J. K.; Holmes, J. G.; and Zanna, M. P. 1985. Trust in close relationships. *Journal of personality and social psychology* 49(1):95.

Setter, T.; Gasparri, A.; and Egerstedt, M. 2016. Trust-based interactions in teams of mobile agents. In *American Control Conference*, 6158–6163.

Steinfeld, A.; Fong, T.; Kaber, D.; Lewis, M.; Scholtz, J.; Schultz, A.; and Goodrich, M. 2006. Common metrics for human-robot interaction. In *Conference on Human-robot interaction*, 33–40. ACM.

Sweet, N.; Ahmed, N. R.; Kuter, U.; and Miller, C. 2016. Towards self-confidence in autonomous systems. In *AIAA Infotech@ Aerospace*. 1651–1652.

Tan, H. H., and Tan, C. S. 2000. Toward the differentiation of trust in supervisor and trust in organization. *Genetic, Social, and General Psychology Monographs* 126(2):241.

van Maanen, P.-P., and van Dongen, K. 2005. Towards task allocation decision support by means of cognitive modeling of trust. In *Int. Workshop on Trust in Agent Societies*, 399–400.