TRUST IN ROBOTS: CHALLENGES AND OPPORTUNITIES

Bing Cai Kok
Department of Computer Science
National University of Singapore
kokbc@comp.nus.edu.sg

Harold Soh
Department of Computer Science
National University of Singapore
harold@comp.nus.edu.sg

ABSTRACT

Purpose of Review: To assess the state-of-the-art in research on trust in robots and to examine if recent methodological advances can aid in the development of trustworthy robots.

Recent Findings: While traditional work in trustworthy robotics has focused on studying the antecedents and consequences of trust in robots, recent work has gravitated towards the development of strategies for robots to actively gain, calibrate and maintain the human user's trust. Among these works, there is emphasis on endowing robotic agents with reasoning capabilities (e.g., via probabilistic modeling).

Summary: The state-of-the-art in trust research provides roboticists with a large trove of tools to develop trustworthy robots. However, challenges remain when it comes to trust in real-world human-robot interaction (HRI) settings: there exist outstanding issues in trust measurement, guarantees on robot behavior (e.g., with respect to user privacy), and handling rich multidimensional data. We examine how recent advances in psychometrics, trustworthy systems, robot-ethics, and deep learning can provide resolution to each of these issues. In conclusion, we are of the opinion that these methodological advances could pave the way for the creation of truly autonomous, trustworthy social robots.

***K*eywords** Trust · Human-Robot Interaction · Probabilistic Models · Privacy · Measurement · Formal Methods

Introduction

On July 2, 1994, USAir Flight 1016 was scheduled to land in the Douglas International Airport in Charlotte, North Carolina. Upon nearing the airport, the plane experienced inclement weather and was affected by wind shear (a sudden change in wind velocity that can destabilize an aircraft). On the ground, a wind-shear alert system installed at the airport issued a total of three warnings to the air traffic controller. But due to a *lack of trust* in the alert system, the air traffic controller transmitted only one of the alarms that was, unfortunately, never received by the plane. Unaware of the presence of wind shear, the aircrew failed to react appropriately and the plane crashed, killing 37 people [1] (see Fig. 1). This tragedy vividly brings to focus the critical role of trust in automation (and by extension, robots): a lack of trust can lead to disuse, with potentially dire consequences. Had the air traffic controller trusted the alert system and transmitted all three warnings, the tragedy may have been averted.

Human-robot trust is crucial in today's world where modern social robots are increasingly being deployed. In healthcare, robots are used for patient rehabilitation [2] and to provide frontline assistance during the on-going COVID-19 pandemic [3,4]. Within education, embodied social robots are being used as tutors to aid learning among children [5]. The unfortunate incident of USAir Flight 1016 highlights the problem of *undertrust*, but maximising a user's trust in a robot may not necessarily lead to positive interaction outcomes. *Overtrust* [6,7] is also highly undesirable, especially in the healthcare and educational settings above. For instance, [6] demonstrated that people were willing to let an unknown robot enter restricted premises (e.g. by holding the door for it), thus raising concerns about security, privacy, and

Figure 1: Remnants of the aircraft N954VJ involved in USAir Flight 1016 [141]. The plane crash was attributed to the lack of warnings about wind shear from the air flight controller on duty who, due to mistrust in the wind shear alert system, reported only one out of three alerts generated by the system.

safety. Perhaps even more dramatically, a study by [7] showed that people willingly ignored the emergency exit sign to follow an evacuation robot taking a wrong turn during a (simulated but realistic) fire emergency, even when said robot performed inefficiently prior to the start of the emergency. These examples drive home a key message: *miscalibrated* trust can lead to *misuse* of robots.

The importance of trust in social robots has certainly not gone unnoticed in the literature — there exist several insightful reviews on trust in robots across a wide spectrum of topics, such as trust repair [8,9], trust in automation [10, 11, 12, 13, 14, 15], trust in healthcare robotics [2], trust measurement [16,17] and probabilistic trust modeling [18]. In contrast to the above reviews, we focus on drawing connections to recent developments in adjacent fields that can be brought to bear upon important outstanding issues in trust research. Specifically, we survey recent advances in the development of trustworthy robots, highlight contemporary challenges, and finally examine how modern tools from psychometrics, formal verification, robot ethics, and deep learning can provide resolution to many of these longstanding problems. Just as how advances in engineering have brought us to the cusp of a robotic revolution, we set out to examine if recent methodological breakthroughs can similarly aid us in answering a fundamentally human question: *do we trust robots to live amongst us and, if not, can we create robots that are able to gain our trust?*

A Question of Trust

To obtain a meaningful answer to our central question, we need to ascertain exactly what is meant by 'trust in a robot'. Is the notion of trust in automated systems (e.g., a wind-shear alert system) equivalent to trust in a social robot? How is trust different from concepts such as trustworthiness and reputation?

Historically, trust has been studied with respect to automated systems [11, 12, 14]. Since then, much effort has been expended to extend the study of trust to human-robot interaction (HRI); see Fig. 2 for an overview. We use the term 'robots' to refer to embodied agents with a physical manifestation that are meant to operate in noisy, dynamic environments [8]. An automated system, on the other hand, may be a computerized process without an explicit physical form. While research on trust in automation can inform our understanding of trust in robots [11], this robot-automation distinction (which is, admittedly, not quite sharp) has important implications for the way we conceptualize trust. For one, the physical embodiment of a robot makes its design a key consideration in the formation of trust. Furthermore, we envision social robots would be typically deployed in dynamic unstructured environments and have to work alongside human agents. This suggests that the ability to navigate uncertainty and social contexts plays a greater role in the formation and maintenance of human trust.

Trust should also be distinguished from the related concepts of trustworthiness and reputation. Trust (in an agent) is a property of the human user in relation to the agent in question [19]. In contrast, trustworthiness is a property of the agent and not of the human user [19, 20]. Hence, a human user may not trust a trustworthy robot (and vice versa); this mismatch is visualised in Fig. 3. The trust-reputation distinction is slightly more nuanced. While both can be thought of as an 'opinion' regarding the agent in question, reputation involves not only the opinion of the single human user (as in trust), but also the collective opinion of other people [21]. In this paper, we mainly focus on human trust in robots.

Distinctions aside, there is, unfortunately, no unified definition of trust in the literature [12,13]. This has led to the proliferation of qualitatively different ways to define trust. For instance, trust has been thought of as a belief [13], an

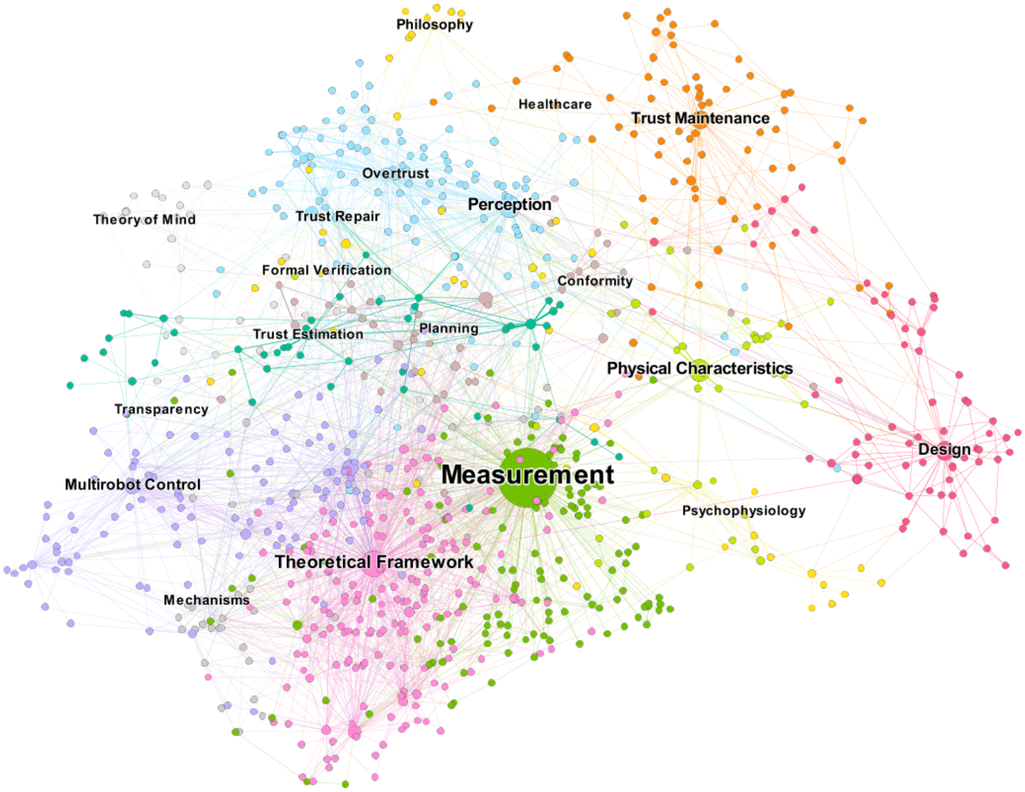

Figure 2: An overview of trust research in robotics. The shown citation network was generated with Gephi [142] based on 925 papers; papers were obtained from an initial list of 72 manually-curated papers on trust in robots. The remaining papers were obtained by crawling the web with OpenCitations [143]. Each node represents a publication, whose size scales with citation count. Initial work concentrated heavily on measuring trust in automation/robots (green node in the middle). Since then, research in the area has branched out to examine areas such as the multirobot control and the design of robots, as well as new theoretical frameworks for understanding trust in robots. The most recent work has explored novel topics such as formal verification in HRI.

attitude [13], an affective response [22], a sense of willingness [23], a form of mutual understanding [24] and as an act of reliance [25].

To handle this ambiguity in definitions, we take a *utilitarian* approach to defining trust for HRI — we adopt a trust definition that gives us practical benefits in terms of developing appropriate robot behavior using planning and control [26,27]. As we will see over the course of this paper, this choice of definition allows us to embed the notion of trust into formal computational frameworks, specifically probabilistic graphical models [28], which in turn allows us leverage powerful computational techniques for estimation, inference, planning and coordination.

We define trust in three parts, each of which is built-off previous work:

– First, the notion of trust can only arise in a situation that involves *uncertainty* and *vulnerability* [23];

– Second, trust is a *multifaceted* construct that is *latent* and cannot be directly observed [12, 29], and

– Third, trust mediates the relationship between the *history of observed events* and the agent's subsequent *act of reliance* [13,19].

Putting everything together, we define an agent's trust in another agent as a

> *'multidimensional latent variable that mediates the relationship between events in the past and the former agent's subsequent choice of relying on the latter in an uncertain environment'*.

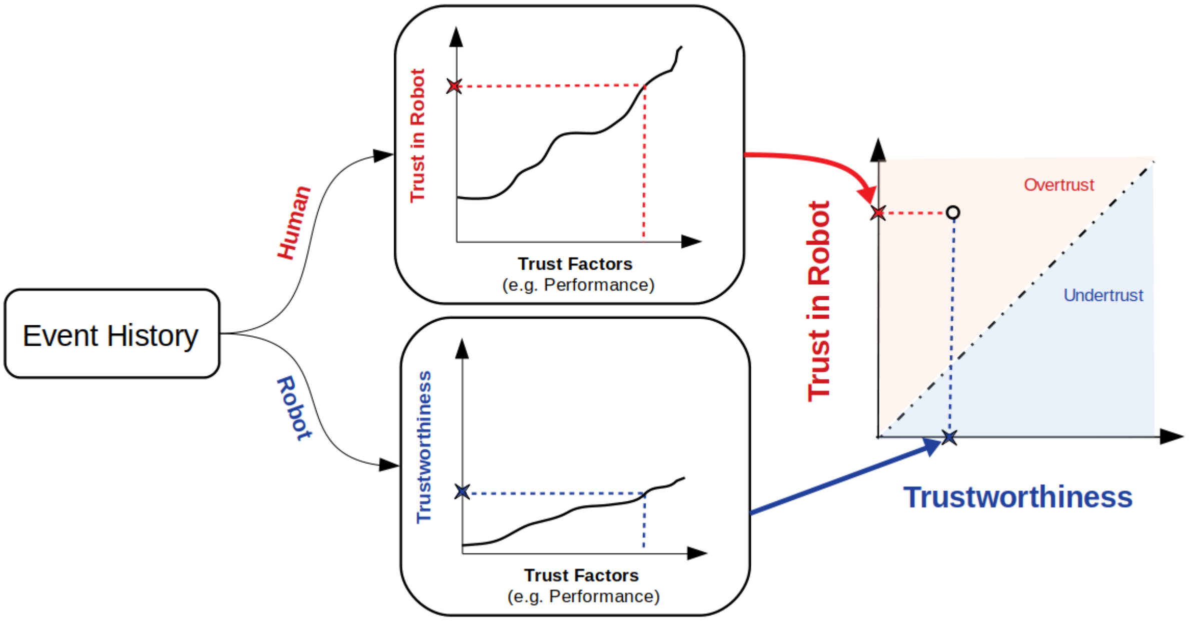

Figure 3: Conceptual diagram of trust calibration. A mismatch between a user's trust in the robot and the robot's actual trustworthiness results in either overtrust or undertrust.

Start with Design

A first step to establish human-robot trust is to design the robot in a way that biases people to trust it appropriately. Conventionally, one way to do so is by configuring the robot's physical appearance [11]. With humans, our decision to trust the person often hinges upon first impressions [30]. The same goes for robots — people often judge a robot's trustworthiness based on its physical appearance [11, 31, 32, 33]. For instance, robots that have human-like characteristics tend to be viewed as more trustworthy [31], but only up to a certain point. Robots that appear to be highly similar, but still quite distinguishable, from humans are seen as less trustworthy [32,33]. This dip in perceived trustworthiness then recovers for robots that appear indistinguishable from humans. This phenomena — called the 'uncanny valley' [8, 32, 33] due to the U-shaped function relating perceived trustworthiness to robot anthropomorphism — has important implications for the design of social robots.

More recently, several works have revealed that the impact of design on trust goes beyond just physical appearance. For example, the way in which the robot is portrayed can greatly impact the perceived trustworthiness of the robot. When a social robot is simply presented as it is, empirical evidence suggests that human users tend to overestimate the robot's capabilities [34] prior to engaging with it. This mismatch between expectation and reality — the *expectation gap* — then leads to an unnecessary loss of trust after users interact with the robot [34]. In a bid to encourage trust formation in the robot, several works have since tried to close this gap by supplying additional information together with the robot. This can be done manually by suitably framing the robot's capabilities prior to any interaction [35, 36, 37]. Alternatively, it can also be done automatically by algorithmically identifying several 'critical states' — those states where the action taken has strong impact on the eventual outcome — from the robot's policy and showing the robot's behavior in those states to the user [38]. These studies demonstrate that suitably engineered supplemental information can help foster human trust in the robot by adjusting people's initial expectations to an appropriate level. In other words, the design of trustworthy robots does not simply stop at considerations about the robot's size, height, and physical make-up. Instead, successful design requires careful thought about the way a robot presents itself to the user.

Gaining, Maintaining, and Calibrating Trust

While a robot's design can prime the user to adopt a certain level of trust in it, the design's efficacy is often dependent on the context [39] and individual differences among human users [40, 41]. As such, design alone may not be sufficient to induce the appropriate level of trust. The robot still has to actively gain, maintain, and calibrate the user's trust for successful collaboration to occur [42]. This is especially challenging considering how meaningful social interactions often occur over a protracted period of time. The user's trust in the robot is not a static phenomenon — trust fluctuates dynamically as the interaction unfolds over time [43]. To cope with this, the robot in question has to deploy effective strategies that enable it to navigate the changing trust landscape. In the following, we have organised existing strategies in the literature into four major groups in order of increasing model complexity, starting with (i) heuristics and (ii) techniques that exploit the interaction process, and progressing to (iii) computational trust models and (iv) Theory of Mind approaches.

Heuristics

An important class of trust calibration strategies in the literature take the form of heuristics: rule-of-thumb responses upon the onset of certain events. The design of these heuristics is often informed by empirical evidence from psychology. Heuristics have been proposed to tackle two different situations: to *combat overtrust* and to *repair trust*. As an example of the former, [25] proposed to use visual prompts to nudge users to reevaluate their trust in the robotic system when the user has left the automated system running unattended for too long.

The second situation occurs when the robot makes a mistake, which can lead to a disproportionate loss of trust in it [44,45]. The potential loss of trust from such events is not just a rare curiosity — robots are far from infallible, especially when operating for prolonged periods of time. It is crucial for the robot to respond appropriately after a mistake so as to reinstate (or recalibrate) the user's trust. That is, it has to engage in *trust repair* [9]. On this front, researchers have documented a variety of relevant repair strategies [8,9]. For example, the robot could provide explanations for the failure or provide alternative plans [9]. However, one should consider the *context* of the situation before selecting a particular repair heuristic. Recent work [46] has shown that when the loss of trust is attributed to the lack of competence in the robot, apologizing can effectively repair trust. In contrast, when the act that induced trust loss was perceived to be intentional on the robot's part, denial that there was any such intention was a better repair strategy compared to an apology. Despite their simplicity, a nuanced application of these heuristics can go a long way in ensuring successful human-robot interaction.

Exploiting the Process

While heuristics focus on *what* a robot should do, a key element for proper trust calibration is to consider *how* the robot behaves. This element can have substantive impact on the user's trust; at the physical level, [47] found that a robot that could convey its incapabilities solely via informative arm movements was found to be more trustworthy by users. Similarly, robots that came across as transparent, either by providing explanations for its actions [48,49] or simply by providing more information about its actions [50, 51], were judged as more trustworthy. At a more cognitive level, one study suggests that robots that took risky actions in an uncertain environment were viewed with distrust [52], although this depends on the individual user's risk appetite [53]. In other work, robots that expressed vulnerability [54] or emotion [55] through natural language were trusted more. Furthermore, the use of language to communicate with the human user has been shown to mitigate the loss of trust that follows from a performance failure [56]. In all of the above, trust can be gained by exploiting the process in which something is accomplished despite the fact that the end goal for the robot does not change.

Computational Trust Models

The techniques reviewed above rely on pre-programmed strategies, which may be difficult to scale for robots that have to operate in multiple different contexts. A more general approach is to directly model the human's dynamic trust in the robot. Work in this area has focused on two problems: (i) estimating trust based on observations of the human's behavior [18, 27, 57, 58, 59, 60, 61, 62, 63, 64] and (ii) utilizing the estimate of trust to guide robot behavior [27, 58, 65, 66, 67, 68, 69].

A major line of work in this area was started with the introduction of Online Probabilistic Trust Inference Model (OPTIMo) [61], which captures trust as a latent variable in a dynamic probabilistic graphical model (PGM) [28]. While there have been other pioneering attempts to model trust, they have been restricted to simple functions [60] or fail to account for uncertainty in the trust estimation [59]. In comparison, the probabilistic graphical approach presented in OPTIMo leverages on powerful inference techniques that allow trust estimation in real time. Furthermore, this approach accounts for both estimation uncertainty and the dynamic nature of trust in a principled fashion by way of Bayesian inference [28]. This approach is perhaps made even more appealing by the evidence in cognitive science that humans act in a Bayes-rational manner [70], suggesting that robots equipped with this variant of trust model is in fact reasoning with a valid approximation of the human user's trust. This graphical model framework also allows us to translate

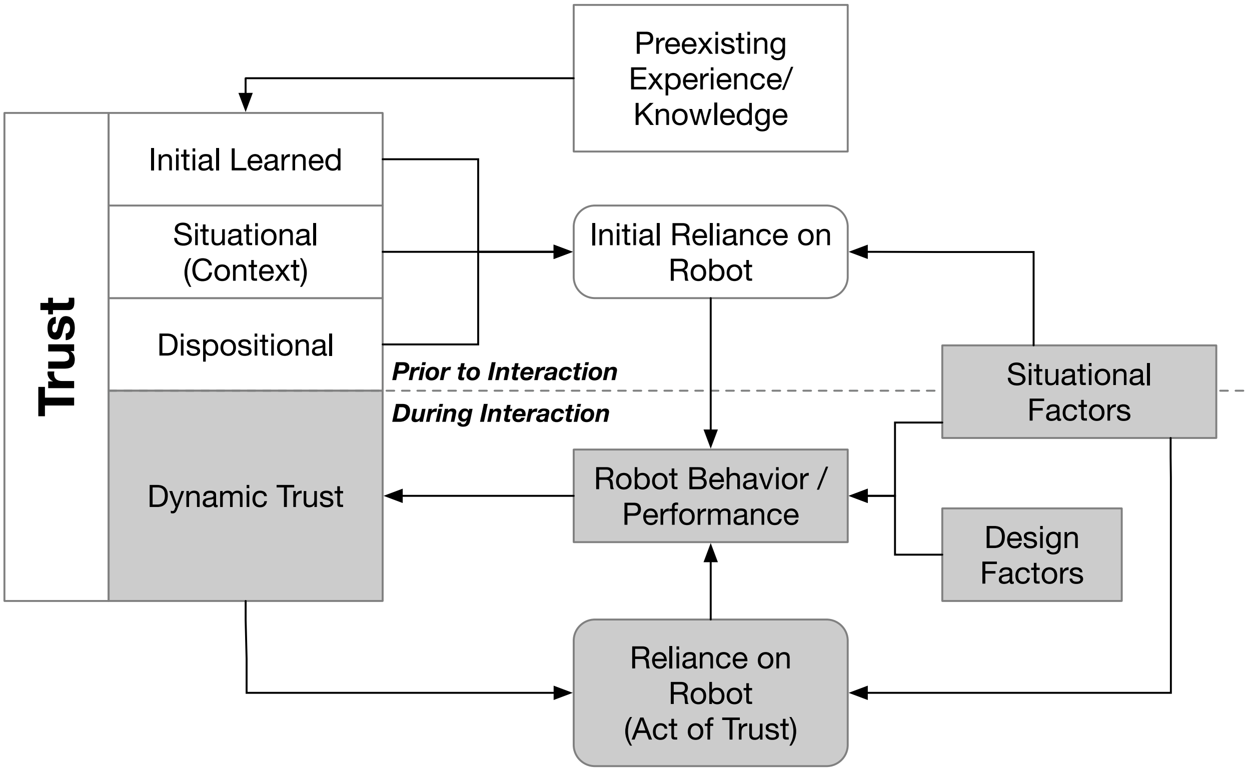

Figure 4: A typical conceptual diagram in the trust literature (e.g., in [12]). Trust is conceptualized as a construct that depends on the user's propensity to trust the robot (i.e. dispositional trust), the task at hand (i.e. situational/contextual trust) and initial 'learned trust' from the user's general prior experience with robots. Trust then shapes the user's initial reliance on the robot. During the interaction, factors can change (grey boxes), which affect robot behavior, and in turn dynamic trust, and the subsequent reliance on the robot.

conceptual diagrams of trust dynamics into an actual quantitative program that lends itself to testable hypotheses [71] (see Fig. 4 and Fig. 5 for an example).

Since the development of OPTIMo, works have contributed important extensions. For example, [64] modelled trust with a beta-binomial (rather than a linear Gaussian) and explored how the model can be used to cluster individuals by their 'trust' profiles. Dynamic Bayesian networks of this flavour has also been used to guide the robot's actions. In several works [27, 68], the estimated trust has been used as a mechanism for the robot to decide if control over the robot should be relinquished to the human. In [67,69], a variant OPTIMo was incorporated into POMDP planning, thus allowing the robot to obtain a policy that reasons over the latent trust of the human user. This Bayesian approach to reasoning about trust has also been explored nonparametrically using Gaussian processes [63]. Lastly, this framework has also been extended to model a user's trust in multiple robots [62], thus paving the way for dynamic trust modeling in multi-agent settings.

Endowing Robots with a Theory of Mind

Trust is but one aspect of the human user's mental state. The most general approach to developing trustworthy robotic agents is to endow them with the ability to reason and respond to the demands and expectations of the human partner in an online, adaptive manner. To do this, recent works have turned to one of the most ubiquitous social agents for inspiration: children.

Despite their nascent mental faculties, children exhibit an astounding ability in navigating our complex social environment. Decades of research has revealed that this social dexterity can be attributed to having a 'Theory of Mind' (ToM) [72]. At a broad level, a ToM refers to an agent's (e.g. a child) ability to mentally represent the beliefs, desires and intentions of another agent (e.g. a different child) [73]. Just as with children, endowing a robot with a ToM would enable it reason about the human user's unobservable mental attributes. Through such reasoning, the robot can predict what the human user wants to do, in turn allowing it to plan in anticipation of the human user's future behavior.

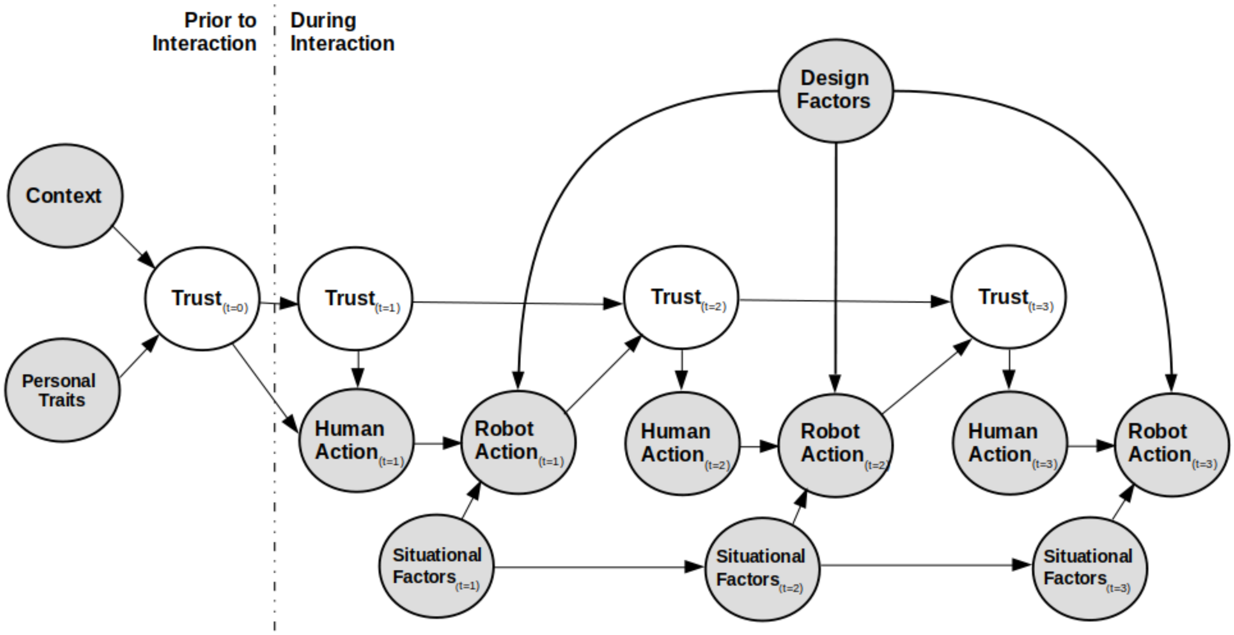

Figure 5: A probabilistic graphical model (PGM) representation of Fig. 4 unrolled for three time steps. Latent variables (i.e. trust in the robot) are shaded. Nodes can represent multidimensional random vectors. The graphical model encodes explicit assumptions about the generative process behind the interaction: conditional independence between any two sets of variables can be inferred by inspecting the structure of the graph via d-separation [28]. By instantiating the prior and conditional probability distributions, the model can be used for inference and simulation.

This idea has been instantiated in robotics before in the formalism behind Interactive Partially Observable Markov Decision Processes [74]. However, it is only recently that this approach has been explored in relation to trust [67, 75, 76, 77] . Recent approaches have shifted away from simple, but unrealistic models of human behavior to empirically motivated ones. For instance, [78] explored a bounded memory model of adaptation where the user was assumed to make decisions myopically, a choice supported by evidence from behavioral economics [79]. Furthermore, they endowed the robot with the ability to reason about the *adaptability* of the user (i.e. the degree to which the user is willing to switch to the robot's preferred plan). By reasoning about this aspect of the human's mental state, the robotic agent could then guide the user towards a strategy that is known to lead to a better outcome when the user is adaptable. In contrast, when the user is adamant about following his/her own policy, then the robot can adapt to the user in order to retain trust. Similarly, other recent works have further explored how incorporating different aspects of the human's beliefs and intentions, such as notions of fairness [76], risk [75] and capability [77], can lead to gains in trust. Endowing robots with a ToM in this way allows robotic agents to better earn the trust of their human partners. That said, full-fledged ToM models remain difficult to construct and generally incur heavier computational cost relative to 'direct' computational trust models.

Summary

Trust is essential for robots to successfully work with their human counterparts. While it is challenging for a robot to gain, maintain and calibrate trust, research in the past decade has provided us with an entire suite of tools to meet this challenge. A note of caution is in order: although we presented these tools in increasing order of complexity, by no means is the order indicative of the relative importance of these tools. The trove of tools developed in the last decade, from heuristics to computational models, is one important step forward to realising a future where social robots can be trusted to live and work amongst us.

Challenges and Opportunities for Trust Research

Scientific progress invariably leads to both new challenges and new opportunities. Research on trust in robots is no exception. We highlight three key challenges and the corresponding opportunities in trust research that could be the focus of much inquiry in the coming decade.

Challenge I: The Measurement of Trust 'In the Wild'

A human's trust in a robot is an unobservable phenomena. As such, there have been major pioneering efforts in the past decade to develop instruments that measure trust in human-robot interaction, most notably in the form of self-report scales [43, 80, 81, 82, 83, 84, 85]. More recently, there has been a trend to move away from self-report measures towards more 'objective' measures of trust [16, 17, 86, 87, 88]. These include physiological measures such as eye-tracking [89, 90], social-cues extracted from video feed / cameras [91, 92], audio [93, 94, 95], skin response [96, 97] and neural measures [96, 97, 98, 99], as well as play behavior in behavioral economic games [37, 92, 100].

This development is in concordance with the broad recognition in the research community on the need to bring HRI from the confines of labs to real-world settings. In this regard, the 'objective' measures above can be an improvement over self-report scales. For instance, physiological measures can be far less disruptive than periodic self-reports, and can be deployed in real-world environments with appropriate equipment. Many of the above measures can also be obtained in real-time with much higher temporal resolution than self-reports, and can be used to directly inform robot decision-making.

Despite this tremendous effort to better measure trust, some thorny issues remain. Even among the validated scales that have been developed, there is a conspicuous lack of confirmatory testing (e.g. via confirmatory rather than exploratory factor analysis). An exception is the trust scale developed in [82] meant to examine trust in automation (not specific to robots), whose factor structure has been confirmed separately in [101]. Furthermore, the literature is relatively silent on the topic of measurement invariance [102, 103]: we found only a single mention in [101]. Briefly, a scale that exhibits measurement invariance is measuring the same construct (i.e. trust) across different comparison groups or occasions [103]. If invariance is not satisfied, then differences in the observed scores between two experimental groups or two time points, even if statistically significant, may not reflect actual differences in trust. Finally, there is a lack of information regarding the psychometric properties, such as the reliability [104], of 'objective' measures. Despite the positive properties mentioned above, 'objective' measures are not immune from psychometric considerations [105]. Rather, 'objective' measures can be thought of as manifestations of underlying trust (e.g., as observations emitted from the latent variable in a PGM) and should be scrutinized for their reliability and validity just as with self-report questionnaires.

Although these issues do not directly invalidate existing results in the literature, it is still important to assess any existing shortcomings in the existing instruments, and improve upon them if need be, especially in view of the replication crisis plaguing psychological science [106].

Challenge II: Bridging Trustworthy Robots and Human Trust

Thus far, we have explored in detail how human users develop trust in robots. However, the literature on human trust in robots can also be seen in relation to research on *trustworthy systems* (and more recently, *trustworthy AI* [107]), where frameworks and methods have been developed to ensure (or assess if) a given system satisfies desired properties. These methods have their roots in the field of software engineering and have traditionally focused on satisfying metrics based on non-functional properties of the system (e.g. reliability) [108] as well as the quality of service (e.g. empathy of a social robot) [108, 109].

More recently, there is a trend to adopt these techniques to model aspects of the human user. One important class of techniques is formal verification [110], which are powerful tools that enable system designers to provide guarantees of system performance. Recent work has extended formal methods to handle concepts such as fairness [111], privacy [112, 113] and even cognitive load [114]. Formal verification techniques have also been applied to problems in human-automation interaction [115, 116], suggesting that these methods can also be fruitfully applied to HRI.

To date, very few works have explored how formal methods can be applied in conjunction with human trust in robots. Pioneering work [117] presented a verification and validation framework that allows assistive robots to demonstrate their trustworthiness in a handover task. Similarly, [118] has explored formal verification methods for cognitive trust in a multiagent setting, in which trust is formalized as an operator in a logic. By combining the logic with a probabilistic model of the environment, one can use formal verification techniques to assess if the particular human-robot system satisfies a list of required properties related to trust. However, much is left to be done. A key technical challenge is to scale-up existing tools to the complexity of HRI. There is also a need to ensure that the models used in formal verification sufficiently represents the HRI task for the guarantees to be meaningful.

One important aspect of trustworthy systems that is intimately related to trust and that deserves additional attention is *privacy*. In [119], the authors raised concern that when social robots are deployed to work with vulnerable groups (e.g. children), there is the possibility that the robots will be used to intrude upon their privacy. For instance, if the vulnerable person develops affection or trust in the robot, they may inappropriately 'confide' in the robot. This can be an avenue for abuse by other malicious agents if the robot has recording capabilities. This idea was further developed in [120] where they showed experimentally that trust in robots can be exploited to convince human users to engage in risky acts (e.g. gambling) and to reveal sensitive information (e.g. information often used in bank password resets). These trust-related privacy issues goes beyond intentional intrusion by external agents when we consider multiagent planning, where there is the possibility of individual information being leaked [121]. This problem becomes all the more acute if medical information (e.g. diagnosis of motor-related diseases, such as Parkinson's) is involved [122]. More broadly, this issue of privacy can be seen from the perspective of developing safe (and thus trustworthy) robots, where the notion of safety includes not only physical safety [123], but also safety from unwanted intrusions of privacy.

Recent work suggests that there are ways to resolve this tension between privacy and trust. In particular, techniques from the differential privacy literature [124] can be used to combat the leakage of private information [122]. With regard to the intrusion of privacy via robots, the field of robot ethics can inform us about the regulations and codes of conduct that should be in place prior to deployment of social robots among vulnerable groups [125]. We should also recognize that the effect of trust on privacy can actually be harnessed for the *social good*. In [126], the development of trust in virtual humans encouraged patients to disclose medical information that they would otherwise have withheld in the presence of a real doctor due to fear of social judgment during a health screening. From this perspective, a robust ethical and legal framework to regulate the use of social robots is important to address privacy and safety concerns associated with the use of social robots in our everyday lives [125].

Challenge III: Rich Trust Models for Real-World Scenarios

Trust is most often described as a rich, multidimensional construct [127] and in a real-world setting, a variety of elements come together to affect trust. While there is no doubt that existing models have yielded impressive results, there is still much work to be done to fully capture trust in robots. Some recent works have begun to explore this area. For example, [53] examined two different aspects of trust — trust in a robot's intention and trust in a robot's capability — and demonstrated that these two factors interact to give rise to reliance on the robot. Similarly, [63] demonstrated that trust can be modelled as a *latent function* (i.e., an infinite-dimensional model of trust) to incorporate different contexts, and showed that this approach could capture how trust transfers across different task environments.

These two examples highlight two distinct but complementary approaches to studying the multidimensional nature of trust. The first is that, by taking multidimensionality seriously, we can develop a richer and more accurate understanding of not just *what* causes humans to cooperate (or not) with robots, but also *how* this cooperation comes about. In other words, it could give us insight into the *mechanisms*, and not just the antecedents, through which trust affects human-robot collaboration. In line with this view, a few pioneering works have begun exploring the mediating role of (unidimensional) trust via formal causal mediation analysis [128, 129, 130]. The goal in this approach is to empirically test if a particular trust-based mechanism is supported by the data [131, 132, 133]. A relatively unexplored, but potentially fruitful, area is the understanding of trust mechanisms in the multidimensional case. There is value in assessing the viability of actual mechanisms beyond scientific curiosity, especially when modelling trust in the PGM framework. In such models, the structure of the graph is often taken to be true by fiat. However, there is no guarantee that the given graph structure corresponds to the actual *causal graph* of the phenomena of interest [28]. Empirical studies that reveal detailed mechanisms underlying trust in robots would be critical in obtaining better approximations to the true causal graph, which in turn should lead to more robust computational models of trust.

A second possible approach is to leverage on the advances in *deep probabilistic models* to model latent trust in the context of *high-dimensional input data*, which is prevalent in real-world settings. This is especially relevant in light of the recent interest in integrating video, audio, and psychophysiological measures into trust modeling [88]. The recent integration of deep neural networks with probabilistic modeling (e.g., [134, 135, 136]) has made it possible to handle high-dimensional unstructured data within PGMs. Briefly, these methods work by mapping the high-dimensional raw data into a reduced space characterized by some latent random real vector, where the mapping is typically achieved via neural networks. Although some may be concerned about the lack of interpretability of the learnt latent space in this approach (due to the nonlinearity of neural networks), recent work has sought to learn latent representations that are '*disentangled*' — in other words, representations whose *individual* dimensions provide meaningful information about the data being modeled [137, 138, 139]. Regardless of interpretability, this approach is nevertheless particularly valuable in the context of robotics: social robots have to make sense of its environment based on raw sensory information. Being able to perform trust inference based on such unstructured data can help a robot to autonomously plan and act appropriately in our social environment.

These two approaches to studying multidimensional trust are not incompatible. The former provides us with the much needed structural understanding of the construct of trust. The latter approach allows robots to effectively handle the rich sources of sensory data to reason about trust in real time. These two approaches can be combined [140] to get the best of both worlds, potentially allowing us to develop robots that can perform structured trust inference in a messy, high-dimensional world.

Conclusions

Human trust in robots is a fascinating and central topic in HRI. Without appropriate trust, robots are vulnerable to either disuse or misuse. In this respect, trust is an enabler that allows robots to emerge from their industrial shell and out into the human social environment. Today, advances in both algorithms and design has led to the creation of social robots that can infer and reason about human characteristics to gain, maintain and calibrate human trust throughout the course of an interaction, further cementing their place as partners to their human users. While challenges remain, methodological advances in recent years seem to promise resolution to these longstanding issues. We trust that the research community will be able to leverage these methods to turn meaningful, long-lasting HRI into an everyday reality.

Compliance with Ethical Standards

Conflict of Interest

The authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

[1] Wald, M.L.: Series of Blunders Led to Crash Of USAir Jet in July, Panel Says (1995)

[2] Langer, A., Feingold-Polak, R., Mueller, O., Kellmeyer,P., Levy-Tzedek, S.: Trust in socially assistive robots: Considerations for use in rehabilitation. Neuroscience and Biobehavioral Reviews104(March), 231–239 (2019). DOI 10.1016/j.neubiorev.2019.07.014

[3] Goh, T.: Coronavirus: Exhibition centre now an isola-tion facility, with robots serving meals (2020). URL https://www.straitstimes.com/singapore/health/exhibition-centre-now-an-isolation-facility-with-robots-serving-meals

[4] Statt, N.: Boston Dynamics' Spot robot is helping hospitals remotely treat coronavirus patients (2020). URL https://www.theverge.com/2020/4/23/21231855/boston-dynamics-spot-robot-covid-19-coronavirus-telemedicine

[5] Belpaeme, T., Kennedy, J., Ramachandran, A., Scas-sellati, B., Tanaka, F.: Social robots for education: A review (2018). DOI 10.1126/scirobotics.aat5954

[6] Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: ACM/IEEE International Conference on Human-Robot Interaction, vol. 2016-April (2016). DOI 10.1109/HRI.2016.7451740

[7] Booth, S., Tompkin, J., Pfister, H., Waldo, J., Gajos,K., Nagpal, R.: Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security. In: ACM/IEEE International Conference on Human-Robot Interaction, vol. Part F1271 (2017). DOI 10.1145 /2909824.3020211

[8] Baker, A.L., Phillips, E.K., Ullman, D., Keebler, J.R.: Toward an understanding of trust repair in human-robot interaction: Current research and future directions. ACM Transactions on Interactive Intelligent Systems 8(4) (2018). DOI 10.1145/3181671

[9] • Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F.,Powers, T.M., Dixon, C., Tielman, M.L.: Taxonomy of trust-relevant failures and mitigation strategies. ACM/IEEE International Conference on Human-Robot Interaction pp. 3–12 (2020). DOI 10.1145 /3319502.3374793

Developed an overarching taxonomy that describes existing approaches to trust-repair in human-robot interaction

[10] Glikson, E., Woolley, A.W.: Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals (April) (2020). DOI 10.5465/annals.2018.0057

[11] Hancock, P.a., Billings, D.R., Schaefer, K.E., Chen,J.Y.C., de Visser, E.J., Parasuraman, R.: A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors 53(5), 517–527 (2011).DOI 10.1177 / 0018720811417254.

[12] Hoff, K.A., Bashir, M.: Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors 57(3) (2015). DOI 10.1177/0018720814547570

[13] Lee, J.D., See, K.A.: Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 46 (1), 50–80 (2004)

[14] Schaefer, K.E., Chen, J.Y., Szalma, J.L., Hancock, P.A.: A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. Human Factors58(3) (2016). DOI 10.1177/0018720816634228

[15] Shahrdar, S., Menezes, L., Nojoumian, M.: A survey on trust in autonomous systems. In: Advances in Intelligent Systems and Computing, vol. 857 (2019). DOI 10.1007/978-3-030-01177-2\27

[16] Basu, C., Singhal, M.: Trust dynamics in human autonomous vehicle interaction: A review of trust models. In: AAAI Spring Symposium - Technical Report, vol.SS-16-01 – (2016)

[17] Brzowski, M., Nathan-Roberts, D.: Trust Measurement in Human–Automation Interaction: A Systematic Review. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63 (1) (2019). DOI10.1177/1071181319631462

[18] Liu, B.: A Survey on Trust Modeling from a Bayesian Perspective (2020). DOI 10.1007/s11277-020-07097-5

[19] Castelfranchi, C., Falcone, R.: Trust Theory (2007)

[20] Ashraf, N., Bohnet, I., Piankov, N.: Decomposing trust and trustworthiness. Experimental Economics 9 (3) (2006). DOI 10.1007/s10683-006-9122-4

[21] Teacy, W.T., Patel, J., Jennings, N.R., Luck, M.: TRAVOS: Trust and reputation in the context of in-accurate information sources. In: Autonomous Agents and MultiAgent Systems, vol. 12 (2006). DOI 10.1007/s10458-006-5952-x

[22] Coeckelbergh, M.: Can we trust robots? Ethics and Information Technology 14 (1) (2012). DOI 10.1007/s10676-011-9279-1

[23] Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organization trust. Academy of Management Review 20 (3) (1995). DOI10.5465/amr.1995.9508080335

[24] Azevedo, C.R., Raizer, K., Souza, R.: A vision for human-machine mutual understanding, trust establishment, and collaboration. In: 2017 IEEE Conference on Cognitive and Computational Aspects of SituationManagement, CogSIMA 2017 (2017). DOI 10.1109/COGSIMA.2017.7929606

[25] Okamura, K., Yamada, S.: Adaptive trust calibration for human-AI collaboration. PLoS ONE 15 (2) (2020).DOI 10.1371/journal.pone.0229132

[26] Chen, M., Nikolaidis, S., Soh, H., Hsu, D., Srinivasa, S.:Trust-Aware Decision Making for Human-Robot Collaboration. ACM Transactions on Human-Robot Interaction 9 (2) (2020). DOI 10.1145/3359616

[27] Wang, Y., Humphrey, L.R., Liao, Z., Zheng, H.: Trust-based multi-robot symbolic motion planning with a human-in-the-loop. ACM Transactions on Interactive Intelligent Systems 8 (4) (2018). DOI 10.1145/321301

[28] Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques (Adaptive Computation and Machine Learning series), vol. 2009 (2009). DOI10.1016/j.ccl.2010.07.006

[29] Lewis, M., Sycara, K., Walker, P.: The Role of Trust in Human-Robot Interaction. In: Studies in Systems, Decision and Control, vol. 117, pp. 135–159. Springer International Publishing (2018). DOI 10.1007/978-3-319-64816-3\8

[30] Yu, M., Saleem, M., Gonzalez, C.: Developing trust: First impressions and experience. Journal of Economic Psychology 43 (2014). DOI 10.1016/j.joep.2014.04.004

[31] Natarajan, M., Gombolay, M.: Effects of anthropomorphism and accountability on trust in human robot interaction. ACM/IEEE International Conference on Human-Robot Interaction pp. 33–42 (2020). DOI 10.1145/3319502.3374839

[32] Z lotowski, J., Sumioka, H., Nishio, S., Glas, D.F., Bart-neck, C.: Appearance of a Robot Affects the Impact of its Behaviour on Perceived Trustworthiness and Empathy pp. 55–66 (2016). DOI 10.1515/pjbr-2016-0005

[33] Mathur, M.B., Reichling, D.B.: Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. Cognition 146 (2016). DOI10.1016/j.cognition.2015.09.008

[34] Kwon, M., Jung, M.F., Knepper, R.A.: Human expectations of social robots. ACM/IEEE International Conference on Human-Robot Interaction 2016 April, 463–464 (2016). DOI 10.1109/HRI.2016.7451807

[35] Xu, J., Howard, A.: The Impact of First Impressions on Human- Robot Trust during Problem-Solving Scenarios. RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication pp.435–441 (2018). DOI 10.1109/ROMAN.2018.8525669

[36] Washburn, A., Adeleye, A., An, T., Riek, L.D.: Robot Errors in Proximate HRI : How Functionality Framing Affects Perceived Reliability and Trust. ACM Transactions on Human Robot Interaction 1 (1), 1–22 (2020)

[37] Law, T., Chita-Tegmark, M., Scheutz, M.: The Interplay Between Emotional Intelligence, Trust, and Gender in Human–Robot Interaction: A Vignette-Based Study. International Journal of Social Robotics (2020). DOI10.1007/s12369-020-00624-1

[38] Huang, S.H., Bhatia, K., Abbeel, P., Dragan, A.D.: Establishing (Appropriate) Trust via Critical States. HRI 2018 Workshop: Explainable Robotic Systems (2018)

[39] Bryant, D., Borenstein, J., Howard, A.: Why should we gender? The effect of robot gendering and occupational stereotypes on human trust and perceived competency. ACM/IEEE International Conference on Human-Robot Interaction pp. 13–21 (2020).DOI 10.1145/3319502.3374778

[40] Bernotat, J., Eyssel, F., Sachse, J.: The (Fe)male Robot: How Robot Body Shape Impacts First Impressions and Trust Towards Robots. International Journal of Social Robotics (2019). DOI 10.1007/s12369-019-00562-7

[41] Agrigoroaie, Roxana, Stefan-Dan Ciocirlan, and Adriana Tapus. "In the Wild HRI Scenario: Influence of Regulatory Focus Theory." Frontiers in Robotics and AI 7 (2020). DOI 10.3389/frobt.2020.00058

[42] De Graaf, M.M., Allouch, S.B., Klamer, T.: Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. Computers in Human Behavior 43 (2015). DOI 10.1016/j.chb.2014.10.030

[43] Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A.,Yanco, H.: Impact of robot failures and feedback on real-time trust. In: ACM/IEEE International Conference on Human-Robot Interaction (2013). DOI10.1109/HRI.2013.6483596

[44] Salomons, N., Van Der Linden, M., Strohkorb Sebo,S., Scassellati, B.: Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. In: ACM/IEEE International Conference on Human-Robot Interaction (2018). DOI 10.1145/3171221.3171282

[45] Salem, M., Lakatos, G., Amirabdollahian, F., Dauten-hahn, K.: 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, HRI '15, pp. 141–148. Association for Computing Machinery, New York, NY, USA (2015). DOI 10.1145/2696454.2696497.

[46] Sebo, S.S., Krishnamurthi, P., Scassellati, B.: 'I Don't Believe You': Investigating the Effects of Robot Trust Violation and Repair. In: ACM/IEEE International Conference on Human-Robot Interaction, vol. 2019-March (2019). DOI 10.1109/HRI.2019.8673169

[47] Kwon, M., Huang, S.H., Dragan, A.D.: Expressing Robot Incapability. In: ACM/IEEE International Conference on Human-Robot Interaction (2018). DOI 10.1145/3171221.3171276

[48] Yang, X.J., Unhelkar, V.V., Li, K., Shah, J.A.: Evaluating Effects of User Experience and System Transparency on Trust in Automation. In: ACM/IEEE International Conference on Human-Robot Interaction, pp. 408–416 (2017)

[49] Wang, N., Pynadath, D.V., Hill, S.G.: Trust calibration within a human-robot team: Comparing automatically generated explanations. ACM/IEEE International Conference on Human-Robot Interaction 2016-April, 109–116 (2016). DOI 10.1109/HRI.2016.7451741

[50] Chen, J.Y., Barnes, M.J.: Agent Transparency for Human-Agent Teaming Effectiveness. In: Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015, pp. 1381–1385. IEEE (2016). DOI 10.1109/SMC.2015.245

[51] Hussein, A., Elsawah, S., Abbass, H.A.: The reliability and transparency bases of trust in human-swarm interaction : principles and implications. Ergonomics 0 (0), 1–17 (2020). DOI 10.1080/00140139.2020.1764112.

[52] Bridgwater, T., Giuliani, M., Van Maris, A., Baker, G.,Winfield, A., Pipe, T.: Examining profiles for robotic risk assessment: Does a robot's approach to risk affect user trust? ACM/IEEE International Conference on Human-Robot Interaction (2), 23–31 (2020). DOI10.1145/3319502.3374804

[53] Xie, Y., Bodala, I.P., Ong, D.C., Hsu, D., Soh, H.:Robot Capability and Intention in Trust-Based Decisions Across Tasks. In: ACM/IEEE International Conference on Human-Robot Interaction, vol. 2019-March (2019). DOI 10.1109/HRI.2019.8673084

[54] Martelaro, N., Nneji, V.C., Ju, W., Hinds, P.: Tell me more: Designing HRI to encourage more trust, disclosure, and companionship. In: ACM/IEEE International Conference on Human-Robot Interaction, vol. 2016-April (2016). DOI 10.1109/HRI.2016.7451750

[55] Hamacher, A., Bianchi-Berthouze, N., Pipe, A.G., Eder,K.: Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. In: 25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016 (2016). DOI10.1109/ROMAN.2016.7745163

[56] Ciocirlan, Stefan-Dan, Roxana Agrigoroaie, and Adriana Tapus: Human-Robot Team: Effects of Communication in Analyzing Trust. In: 28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019 (2019).

[57] Nam, C., Walker, P., Li, H., Lewis, M., Sycara,K.: Models of Trust in Human Control of Swarms With Varied Levels of Autonomy. IEEE Transactions on Human-Machine Systems (2019).DOI 10.1109/THMS.2019.2896845

[58] Xu, A., Dudek, G.: Trust-driven interactive visual navigation for autonomous robots. In: Proceedings - IEEE International Conference on Robotics and Automation, pp. 3922–3929. Institute of Electrical and Electronics Engineers Inc. (2012). DOI 10.1109 /ICRA.2012.6225171

[59] Akash, K., Hu, W.L., Reid, T., Jain, N.: Dynamic modeling of trust in human-machine interactions. Proceedings of the American Control Conference pp. 1542–1548(2017). DOI 10.23919/ACC.2017.7963172

[60] Floyd, M.W., Drinkwater, M., Aha, D.W.: Adapting autonomous behavior using an inverse trust estimation. In: Lecture Notes in Computer Science (including sub-series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8579 LNCS (2014).DOI 10.1007/978-3-319-09144-0\50

[61] Xu, A., Dudek, G.: OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15 pp. 221–228 (2015).DOI10.1145/2696454.2696492.

[62] Zheng, H., Liao, Z., Wang, Y.: Human-robot trust integrated task allocation and symbolic motion planning for heterogeneous multi-robot systems. In: ASME 2018 Dynamic Systems and Control Conference, DSCC 2018,vol. 3 (2018). DOI 10.1115/DSCC2018-9161

[63] Soh, H., Xie, Y., Chen, M., Hsu, D.: Multi-task trust transfer for human–robot interaction. International Journal of Robotics Research39(2-3) (2020). DOI 10.1177/0278364919866905

[64] Guo, Y. Zhang, C., Yang, X. J.: Modeling Trust Dynamics in Human-robot Teaming : A Bayesian Inference Approach. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems pp. 1–7 (2020) DOI 10.1145/3334480.3383007

[65] Liu, R., Jia, F., Luo, W., Chandarana, M., Nam, C.,Lewis, M., Sycara, K.: Trust-aware behavior reflection for robot swarm self-healing. In: Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, vol. 1 (2019)

[66] Liu, R., Cai, Z., Lewis, M., Lyons, J., Sycara, K.: Trust Repair in Human-Swarm Teams+. In: 2019 28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019 (2019). DOI10.1109/RO-MAN46459.2019.8956420

[67] Chen, M., Nikolaidis, S., Soh, H., Hsu, D., Srinivasa, S.: Planning with Trust in Human Robot Collaboration.In: ACM/IEEE International Conference on Human-Robot Interaction, pp. 307–315. ACM, Chicago, IL,USA (2018)

[68] Xu, A., Dudek, G.: Maintaining efficient collaboration with trust-seeking robots. IEEE International Conference on Intelligent Robots and Systems 2016-Novem,3312–3319 (2016). DOI 10.1109/IROS.2016.7759510

[69] • Chen, M., Nikolaidis, S., Soh, H., Hsu, D., Srinivasa, S.: Trust-Aware Decision Making for Human-Robot Collaboration. ACM Transactions on Human-Robot Interaction 9(2) (2020). DOI 10.1145/3359616

Demonstrated how computational trust models based on PGMs can be embedded into a robot planning framework to allow for smoother human-robot interaction.

[70] Sanborn, A.N., Chater, N.: Bayesian Brains without Probabilities (2016). DOI 10.1016/j.tics.2016.10.003

[71] Airoldi, E.M.: Getting started in probabilistic graphical models (2007). DOI 10.1371/journal.pcbi.0030252

[72] Doherty, M.J.: Theory of mind: How children under-stand others' thoughts and feelings (2008). DOI 10.4324/9780203929902

[73] Yott, J., Poulin-Dubois, D.: Are Infants' Theory-of-Mind Abilities Well Integrated? Implicit Understanding of Intentions, Desires, and Beliefs. Journal of Cognition and Development 17 (5) (2016). DOI 10.1080/15248372.2015.1086771

[74] Gmytrasiewicz, P.J., Doshi, P.: A framework for sequential planning in multi-agent settings. Journal of Artificial Intelligence Research 24 (2005). DOI 10.1613/jair.1579

[75] Kwon, M., Biyik, E., Talati, A., Bhasin, K., Losey, D.P., Sadigh, D.: When humans aren't optimal: Robots that collaborate with risk-aware humans. In: ACM/IEEEInternational Conference on Human-Robot Interaction (2020). DOI 10.1145/3319502.3374832

[76] Claure, H., Chen, Y., Modi, J., Jung, M., Nikolaidis, S.: Reinforcement Learning with Fairness Constraints for Resource Distribution in Human-Robot Teams (2019).

[77] Lee, J., Fong, J., Kok, B.C., Soh, H.: Getting to know one another: Calibrating intent, capabilities and trust for collaboration. IEEE International Conference on Intelligent Robots and Systems (2020).

[78] Nikolaidis, S., Kuznetsov, A., Hsu, D., Srinivasa, S.: Formalizing human-robot mutual adaptation: A bounded memory model. ACM/IEEE International Conference on Human-Robot Interaction 2016-April, 75–82(2016). DOI 10.1109/HRI.2016.7451736

[79] Yi, R., Gatchalian, K.M., Bickel, W.K.: Discounting of past outcomes. Experimental and Clinical Psychopharmacology 14 (3) (2006). DOI 10.1037/1064-1297.14.3.311

[80] Muir, B.M.: Operators' trust in and use of automatic controllers in a supervisory process control task. Ph.D.thesis, University of Toronto (1989)

[81] Merritt, S.M., Heimbaugh, H., LaChapell, J., Lee, D.: I Trust It, but I Don't Know Why. Human Factors: The Journal of the Human Factors and Ergonomics Society55(3) (2013). DOI 10.1177/0018720812465081

[82] Jian, J.Y., Bisantz, A.M., Drury, C.G.: Foundations for an Empirically Determined Scale of Trust in Automated Systems. International Journal of Cognitive Ergonomics 4 (1) (2000). DOI 10.1207/s15327566ijce0401\04

[83] Schaefer, K.: The Perception And Measurement Of Human-robot Trust (2013).

[84] Schaefer, K.E.: Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In: Robust Intelligence and Trust in Autonomous Systems (2016). DOI 10.1007/978-1-4899-7668-0\10

[85] Korber, M.: Theoretical considerations and development of a questionnaire to measure trust in automation. In: Advances in Intelligent Systems and Computing, vol.823 (2019). DOI 10.1007/978-3-319-96074-6\2

[86] Zhou, J., Chen, F.: DecisionMind: revealing human cognition states in data analytics-driven decision making with a multimodal interface. Journal on Multimodal User Interfaces 12(2) (2018). DOI 10.1007/s12193-017-0249-8

[87] Lucas, G., Stratou, G., Gratch, J., Lieblich, S.: Trust me: Multimodal signals of trustworthiness. In: ICMI2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction (2016). DOI10.1145/2993148.2993178

[88] Nahavandi, S.: Trust in Autonomous Systems-iTrustLab: Future Directions for Analysis of Trust With Autonomous Systems.IEEE Systems, Man, and Cybernetics Magazine5(3) (2019). DOI 10.1109 /msmc.2019.2916239

[89] Jenkins, Q., Jiang, X.: Measuring trust and application of eye tracking in human robotic interaction. In: IIEAnnual Conference and Expo 2010 Proceedings (2010)

[90] Lu, Y., Sarter, N.: Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability. IEEE Transactions on Human-Machine Systems 49(6) (2019). DOI10.1109/THMS.2019.2930980

[91] Lee, J.J., Knox, B., Breazeal, C.: Modeling the Dynamics of Nonverbal Behavior on Interpersonal Trust for Human-Robot Interactions. Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium (2008), 46–47 (2013)

[92] Lee, J.J., Knox, W.B., Wormwood, J.B., Breazeal, C.,DeSteno, D.: Computationally modeling interpersonal trust. Frontiers in Psychology4(DEC), 1–14 (2013). DOI 10.3389/fpsyg.2013.00893

[93] Khalid, H., Liew, W.S., Voong, B.S., Helander, M.: Creativity in Measuring Trust in Human-Robot Interaction Using Interactive Dialogs. In: Advances in Intelligent Systems and Computing, vol. 824 (2019). DOI10.1007/978-3-319-96071-5\119

[94] Khalid, H.M., Shiung, L.W., Nooralishahi, P., Rasool,Z., Helander, M.G., Kiong, L.C., Ai-vyrn, C.: ExploringPsycho-Physiological Correlates to Trust. Proceedings of the Human Factors and Ergonomics Society AnnualMeeting 60 (1) (2016). DOI 10.1177/1541931213601160

[95] Elkins, A.C., Derrick, D.C.: The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents. Group Decision and Negotiation22(5) (2013). DOI 10.1007/s10726-012-9339-x

[96] Akash, K., Hu, W.L., Jain, N., Reid, T.: A classification model for sensing human trust in machines using EEG and GSR. ACM Transactions on Interactive IntelligentSystems8(4) (2018). DOI 10.1145/3132743

[97] Hu, W.L., Akash, K., Jain, N., Reid, T.: Real-Time Sensing of Trust in Human-Machine Interactions. IFAC-PapersOnLine49(32) (2016). DOI 10.1016/j.ifacol.2016.12.188

[98] Ajenaghughrure, I.B., Sousa, S.C., Kosunen, I.J.,Lamas, D.: Predictive Model to Assess User Trust: A Psycho-Physiological Approach. In: Proceedings of the10th Indian Conference on Human-Computer Interaction, IndiaHCI '19. Association for Computing Machinery, New York, NY, USA (2019). DOI 10.1145/3364183.3364195

[99] Gupta, K., Hajika, R., Pai, Y.S., Duenser, A., Lochner,M., Billinghurst, M.: Measuring Human Trust in a Virtual Assistant using Physiological Sensing in Virtual Reality. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 756–765 (2020)

[100] Mota, R.C., Rea, D.J., Le Tran, A., Young, J.E., Sharlin, E., Sousa, M.C.: Playing the 'trust game' with robots: Social strategies and experiences. 25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016 pp. 519–524(2016). DOI 10.1109/ROMAN.2016.7745167

[101] Spain, R.D., Bustamante, E.A., Bliss, J.P.: Towards an empirically developed Scale for System Trust: Take two. In: Proceedings of the Human Factors and Ergonomics Society, vol. 2 (2008). DOI 10.1177 /154193120805201907

[102] Borsboom, D.: The attack of the psychometricians. Psychometrika71(3) (2006). DOI 10.1007/s11336- 006-1447-6

[103] Putnick, D.L., Bornstein, M.H.: Measurement invariance conventions and reporting: The state of the art and future directions for psychological research (2016). DOI 10.1016/j.dr.2016.06.004

[104] Elliott, M., Knodt, A., Ireland, D., Morris, M., Poulton, R., Ramrakha, S., Sison, M., Moffitt, T., Caspi,A., Hariri, A.: What is the Test-Retest Reliability of Common Task-fMRI Measures? New Empirical Evidence and a Meta-Analysis. Biological Psychiatry 87 (9),S132–S133 (2020). DOI 10.1016/j.biopsych.2020.02.356

[105] Burt, K.B., Obradovic, J.: The construct of psychophysiological reactivity: Statistical and psychometric issues (2013). DOI 10.1016/j.dr.2012.10.002

[106] Open Science Collaboration: Estimating the reproducibility of psychological science. Science349(6251)(2015). DOI 10.1126/science.aac4716

[107] High Level Independent Group on Artificial Intelligence (AI HLEG): Ethics Guidelines for Trustworthy AI. Tech. Rep. (2019)

[108] Ramaswamy, A., Monsuez, B., & Tapus, A.: Modeling non-functional properties for human-machine systems. In 2014 AAAI Spring Symposium Series. 2014.

[109] Ramaswamy, A., Monsuez, B., & Tapus, A.: SafeRobots: A model-driven Framework for developing Robotic Systems. IEEE International Conference on Intelligent Robots and Systems, 1517-1524 (2014). DOI 10.1109/IROS.2014.6942757

[110] Michael, J.B., Drusinsky, D., Otani, T.W., Shing, M.T.:Verification and validation for trustworthy software systems. IEEE Software 28(6) (2011). DOI 10.1109/MS.2011.151

[111] Si, Y., Sun, J., Liu, Y., Dong, J.S., Pang, J., Zhang, S.J.,Yang, X.: Model checking with fairness assumptions using PAT. Frontiers of Computer Science 8(1) (2014). DOI 10.1007/s11704-013-3091-5

[112] Tschantz, M.C., Kaynar, D., Datta, A.: Formal verification of differential privacy for interactive systems (extended abstract). In: Electronic Notes in Theoretical Computer Science, vol. 276 (2011). DOI 10.1016/j.entcs.2011.09.015

[113] Joshaghani, R., Sherman, E., Black, S., Mehrpouyan,H.: Formal specification and verification of user-centric privacy policies for ubiquitous systems. In: ACM International Conference Proceeding Series (2019). DOI 10.1145/3331076.3331105

[114] Ruksenas, R., Back, J., Curzon, P., Blandford, A.:Verification-guided modelling of salience and cognitive load. Formal Aspects of Computing 21(6) (2009). DOI10.1007/s00165-008-0102-7

[115] Curzon, P., Ruksenas, R., Blandford, A.: An approach to formal verification of human-computer interaction. Formal Aspects of Computing19(4) (2007). DOI 10.1007/s00165-007-0035-6

[116] Bolton, M.L., Bass, E.J., Siminiceanu, R.I.: Using formal verification to evaluate human-automation interaction: A review. IEEE Transactions on Systems, Man,and Cybernetics Part A: Systems and Humans 43(3) (2013). DOI 10.1109/TSMCA.2012.2210406

[117] Webster, M., Western, D., Araiza-Illan, D., Dixon, C.,Eder, K., Fisher, M., Pipe, A.G.: A corroborative approach to verification and validation of human–robot teams. International Journal of Robotics Research 39(1) (2020). DOI 10.1177/0278364919883338

[118] • Huang, X., Kwiatkowska, M., Olejnik, M.: Reasoning about Cognitive Trust in Stochastic Multiagent Systems. ACM Transactions on Computational Logic 20(4) (2019). DOI 10.1145/3329123

Explored how trust can be formulated as an operator in a logic, thereby bringing techniques from formal verification into the study of cognitive trust in multiagent systems.

[119] Sharkey, A.J.: Should we welcome robot teachers? Ethics and Information Technology 18(4) (2016). DOI 10.1007/s10676-016-9387-z

[120] Aroyo, A.M., Rea, F., Sandini, G., Sciutti, A.: Trust and Social Engineering in Human Robot Interaction: Will a Robot Make You Disclose Sensitive Information, Conform to Its Recommendations or Gamble? IEEE Robotics and Automation Letters 3(4) (2018). DOI10.1109/LRA.2018.2856272

[121] Stolba, M., Tozicka, J., Komenda, A.: Quantifying privacy leakage in multi-agent planning. ACM Trans-actions on Internet Technology 18(3) (2018). DOI 10.1145/3133326

[122] Given-Wilson, T., Legay, A., Sedwards, S.: Information security, privacy, and trust in social robotic assistants for older adults. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10292LNCS (2017). DOI 10.1007/978-3-319-58460-7\7

[123] Maurtua, I., Ibarguren, A., Kildal, J., Susperregi, L., & Sierra, B. (2017). Human–robot collaboration in industrial applications: Safety, interaction and trust. International Journal of Advanced Robotic Systems,14(4), DOI 10.1177/1729881417716010

[124] Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9(3-4) (2013). DOI 10.1561/0400000042

[125] Winfield, A.F., Jirotka, M.: Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376(2133) (2018). DOI 10.1098/rsta.2018.0085

[126] Lucas, G.M., Gratch, J., King, A., Morency, L.P.: It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior 37 (2014). DOI 10.1016/j.chb.2014.04.043

[127] Lewis, M., Sycara, K., Walker, P.: The Role of Trust in Human-Robot Interaction. In: Studies in Systems, Decision and Control, vol. 117, pp. 135–159. Springer International Publishing (2018). DOI 10.1007/978-3-319-64816-3\8

[128] VanderWeele, T.: Explanation in causal inference: methods for mediation and interaction. Oxford University Press (2015)

[129] Gonzalez, O., MacKinnon, D.P.: The Measurement of the Mediator and Its Influence on Statistical Mediation Conclusions. Psychological Methods (2020). DOI10.1037/met0000263

[130] Muthen, B., Asparouhov, T.: Causal Effects in Mediation Modeling: An Introduction With Applications to Latent Variables. Structural Equation Modeling 22(1)(2015). DOI 10.1080/10705511.2014.935843

[131] Hussein, A., Elsawah, S., Abbass, H.A.: Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. Human Factors (2019). DOI 10.1177/0018720819879273

[132] Chancey, E.T., Bliss, J.P., Proaps, A.B., Madhavan, P.: The Role of Trust as a Mediator between System Characteristics and Response Behaviors. Human Factors 57(6) (2015). DOI 10.1177/0018720815582261

[133] Bustamante, E.A.: A reexamination of the mediating effect of trust among alarm systems' characteristics and human compliance and reliance. In: Proceedings of theHuman Factors and Ergonomics Society, vol. 1 (2009). DOI 10.1518/107118109x12524441080344

[134] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes (VAE, reparameterization trick). ICLR 2014 (Ml) (2014)

[135] Krishnan, R.G., Shalit, U., Sontag, D.: Structured inference networks for nonlinear state space models. In: 31st AAAI Conference on Artificial Intelligence, AAAI 2017 (2017)

[136] Tan, Z.X., Soh, H., Ong, D.: Factorized Inference in Deep Markov Models for Incomplete Multimodal Time Series. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34

[137] Ansari, A.F., Soh, H.: Hyperprior Induced Unsupervised Disentanglement of Latent Representations. Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019). DOI 10.1609/aaai.v33i01.33013175

[138] Hsu, W.N., Zhang, Y., Glass, J.: Unsupervised learning of disentangled and interpretable representations from sequential data. In: Advances in Neural Information Processing Systems, vol. 2017-Decem (2017)

[139] Li, Y., Mandt, S.: Disentangled sequential autoencoder. In: 35th International Conference on Machine Learning, ICML 2018, vol. 13 (2018)

[140] Johnson, M.J., Duvenaud, D., Wiltschko, A.B., Datta,S.R., Adams, R.P.: Composing graphical models with neural networks for structured representations and fast inference. In: Advances in Neural Information Processing Systems (2016)

[141] Wikimedia Commons, the free media repository: File:Usairflight1016(4).jpg (2017). URL https://commons.wikimedia.org/w/index.php?title=File:USAirFlight1016(4).jpg&oldid=261398935. [On-line; accessed 16-June-2020]

[142] Bastian, M., Heymann, S., Jacomy, M.: Gephi: Anopen source software for exploring and manipulatingnetworks (2009).

[143] Peroni, S., Shotton, D.: OpenCitations, an infrastruc-ture organization for open scholarship. Quantitative Sci-ence Studies1(1) (2020). DOI 10.1162/qss\a\00023