CHAPTER

15

Trust-aware control in proximate human-robot teaming

Auriel Washburn, Sachiko Matsumoto, and Laurel D. Riek

Computer Science and Engineering, UC San Diego, La Jolla, CA, United States

Introduction

Proximate interaction is central to a wide variety of rapidly expanding human-robot interaction (HRI) domains including healthcare, manufacturing, and education (see Fig. 1). Recent findings from HRI demonstrate that human-robot teaming commonly leads to better performance than humans achieve alone (e.g., de Visser & Parasuraman, 2011; Dixon & Wickens, 2006; Marble, Bruemmer, Few, & Dudenhoeffer, 2004; McKendrick et al., 2014; Shaw et al., 2010). This work suggests that proximate human-robot teaming can support notable improvements in many aspects of everyday life.

A primary goal of HRI research in general is to establish how to achieve and maintain trust calibration within a human-robot team (Ogreten, Lackey, & Nicholson, 2010). Trust calibration is the achievement of an appropriate level of trust given both human and robot capabilities and serves to support effective cooperation and reliance between agents. This prevents situations where humans over- or under-rely on their robot teammates, which can result in ineffective robot monitoring or a lack of interaction with the robot (Hancock et al., 2011). Human-robot trust calibration is critical to effective proximate human-robot teaming, as such teaming often occurs in dynamic environments within which human actions are highly variable and robot errors are likely to occur including public spaces like hospitals, schools and airports as well as private homes. The continued development of trust-aware robot control will greatly

https://doi.org/10.1016/B978-0-12-819472-0.00015-0

FIG. 1 Examples of robots engaging in proximate HRI from our prior work (clockwise from top left): collaboratively hanging a banner (Washburn, Adeleye, An, & Riek, 2020), engaging in rehabilitation exercises (Woodworth, Ferrari, Zosa, & Riek, 2018), stacking boxes (Matsumoto, Washburn, & Riek, 2018), and implicitly learning from a human (Hayes, Moosaei, & Riek, 2016). These platforms act in shared spaces with people, illustrating the primary importance of human safety and robot movement to trust in proximate HRI.

support the maintenance of trust calibration during proximate human-robot teaming.

This chapter provides the foundation for a new generation of trust-aware control frameworks specifically designed to support optimal teaming during proximate HRI. To achieve this we review existing work across HRI, human-automation interaction, and human-human interaction (HHI) and highlight opportunities for the advancement of trust-aware control in proximate HRI, as well current challenges to this progress.

In the "Background" section, we provide background on the modeling of trust in HRI. In "Critical trust factors within proximate HRI", we review the interaction factors that are most likely to affect trust during proximate HRI. We then propose a set of five aims for trust-aware control in proximate HRI in "Proposed aims for trust-aware control in proximate HRI". In "Existing trust-aware control frameworks for HRI", we use these aims to

assess the advantages and shortcomings of two existing human-robot trust-aware control frameworks with respect to their use in proximate HRI. Through this process, robot movement behavior emerges as central to trust dynamics in proximate human-robot teaming, and in "Advancing trust-aware control of robot movement in proximate HRI", we offer suggestions for the further advancement of trust-aware modeling for proximate human-robot teaming through knowledge about robot movement behavior. In "Challenges to trust-aware control in proximate HRI", we describe continued challenges in achieving trust-aware control within proximate HRI. Lastly, in "The future of proximate HRI", we summarize our recommended priorities for the development of trust-aware control frameworks designed to optimize proximate human-robot teaming.

Background

Theoretical framework for understanding trust in HRI

The current theoretical framework for understanding trust in HRI is largely informed by earlier work on trust in human-automation interaction. Lee and See (2004) articulated one commonly accepted understanding of trust in HRI with respect to trust in human-automation interaction as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". Many HRI researchers also agree with the assertion by Mayer, Davis, and Schoorman (1995) that human trust in an automated agent involves recognizing and accepting risk.

Beyond these principle elements, empirical work on trust in HRI reveals various factors that impact trust. Hancock et al. (2011) conducted a comprehensive review of this work and found that these factors generally fall into three categories: 1) robot-related, 2) human-related, and 3) environmental. The majority of research on trust in HRI has focused on robot-related factors, which can be divided into attribute- or performance-based robot characteristics (Hancock et al., 2011). Attribute-based characteristics include the way a robot appears, whereas performance-based characteristics focus on the way the robot functions. Attribute-based traits related to trust include robot type (e.g., fixed arm, mobile manipulator, or autonomous vehicle), personality, and anthropomorphism. For example, robots with polite personalities can elicit trust even when they exhibit low reliability (Parasuraman & Miller, 2004). Performance-based features, including robot reliability (De Brun et al., 2008; Lee & See, 2004; Ogreten et al., 2010) and predictability (de Vries, Midden, & Bouwhuis, 2003; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Lee & See, 2004), are consistently and

positively associated with human attributions of trust. Hancock et al. (2011) identified performance-based robot characteristics as prominent predictors of human trust during HRI. Among these characteristics, robot reliability was the most consistent predictor of human trust across studies.

However, Ososky, Schuster, Phillips, and Jentsch (2013) suggested that trust calibration is more a function of an individual's understanding of a robot's abilities and inabilities rather than the ground-truth performance of the robot. Prior work suggests that individuals with low expectations for a robot's capabilities have more positive experiences during HRI (Paepcke & Takayama, 2010). Our recent work demonstrated that whether or not an individual expects a robot to make errors significantly affects their experience of trust during interaction with a robot that actually makes errors (Washburn et al., 2020). Specifically, individuals with low expectations reported greater trust recovery following a robot error than those with high expectations.

Both Paepcke and Takayama (2010) and Washburn et al. (2020) gave their participants framing information about robot capabilities prior to interaction, but expectations for robot behavior can also be shaped by robot morphology. For example, people often attribute a range of social abilities to humanoid robots and can be disappointed and dissatisfied when a robot does not exhibit these skills during interaction (Duffy, 2003; Mori, 1970). It is, therefore, important to keep in mind that some attribute-based, robot-related factors likely impact trust during HRI indirectly by creating a mismatch between human expectations and robot abilities.

Work in human factors also suggests a robust relationship between self-confidence and trust (Freedy, DeVisser, Weltman, & Coeyman, 2007; Lee & Moray, 1994; Ogreten et al., 2010). Specifically, individuals with low self-confidence are more likely to overtrust a robot while those with high self-confidence are less likely to achieve trust at all. People with a high propensity to trust other people or things, in general, are more likely to develop appropriate trust in automation (Lee & See, 2004). This may also be applied to how humans develop trust in robots (Adams, Bruyn, Houde, & Angelopoulos, 2003).

Compared to the number of studies focused on identifying the role of robot-related and human-related factors in human-robot trust, considerably fewer studies investigate the impact of environment-related factors. At the time of their review, Hancock et al. (2011) identified two studies that demonstrated an effect of cultural context on trust ratings for robots (Lee & See, 2004; Li, Rau, & Li, 2010). They also predicted that various elements of team collaboration (e.g., communication) and task features like the task type or need for multitasking would influence trust during human-robot teaming.

Modeling trust in human-automation teaming

In their previous work, Yanco, Desai, Drury, and Steinfeld (2016) and Moray and Inagaki (1999) identified five distinct categories of trust models for human-automation interaction: qualitative, regression, time series, neural net, and probabilistic models. The authors discuss the pros and cons of the model categories, ultimately demonstrating that each type offers different opportunities for supporting trust and trust calibration during interaction.

Qualitative models consist of relationships and interactions between independent variables of interest (i.e., robot-related, human-related or environmental factors) and trust as a dependent variable (e.g., Muir, 1989; Riley, 1994). *Regression models* use statistical techniques, such as multiple linear regressions, to examine the relationship between the independent variables and the dependent variable, typically trust. They also identify which independent variables account for the most variance in trust levels. *Time series models* leverage the structured nature of time series data to model fluctuations in trust over the course of interaction. Since these models incorporate time, they are able to represent dynamic relationships between independent variables and trust, unlike the regression models described earlier.

Neural net models train a neural net on collected data to predict a human's future behavior or trust level. However, unlike time series models, the process of generating neural nets does not allow the developer to extract meaningful information about how the model works. As a result, a developer may be unaware of the ways in which a neural net model is unintentionally biased by characteristics of the training data set. This may make it challenging to correct errors introduced by the training data or adjust the model for use in related but distinctly different contexts. *Probabilistic models* are meant to approximate high-level human cognitive activity by identifying the probability that a specific agent or robot action will support the team's goal state, as well as the agent's confidence in its decision.

Aside from purely qualitative models, each of the four other model types offers opportunities for direct implementation on robotic platforms to provide online control during interaction. Of these quantitative model types, time series, neural nets, and probabilistic models can all be designed to account for the dynamic nature of trust in HRI (Robinette, Howard, & Wagner, 2015; Rousseau, Sitkin, Burt, & Camerer, 1998; Salem, Lakatos, Amirabdollahian, & Dautenhahn, 2015). Given the need for extensive training and the challenges associated with adjusting neural net models, versions of time series, and probabilistic models are likely to be best-suited to supporting real-time, trust-aware control of robot

behavior. As Yanco et al. (2016) discussed, these models could even allow a robot to engage in corrective behavior during interaction. The continued progress of proximate HRI depends on the development of quantitative control models that account for the factors with the greatest impact on trust within proximate interaction.

Trust in proximate HRI

In proximate HRI, the colocation of the human and robot during teaming shapes the influence of many human-related, robot-related, and environmental factors on trust. Thus, it is important to distinguish copresent interaction, in which a human and robot share the same physical space, from tele-present interaction, in which a human and robot interact with each other from distinct physical locations. Prior work demonstrates that copresent HRI engenders a greater degree of human trust than telepresent interaction with the same robot (Bainbridge, Hart, Kim, & Scassellati, 2008). With the rapid increase of opportunities for proximate interaction within healthcare, manufacturing, and education it is crucial that we understand the features of human-robot trust specific to proximate HRI (Lasota & Shah, 2015; Riek, 2017).

Critical trust factors within proximate HRI

As noted previously, past work in human-automation interaction informs a large proportion of work on trust in HRI across proximate and telepresent contexts. However, Desai, Stubbs, Steinfeld, and Yanco (2009) cautioned that robotics involves an increased level of uncertainty and vulnerability beyond that experienced with automation. For instance, HRI often occurs within unstructured environments. With current robots depending on noisy sensors, the incomplete information available to them in these environments can reduce the reliability of robot behaviors (Yanco et al., 2016). Additionally, while the risk associated with human-automation interaction tasks can vary widely, the majority of these studies focus on scenarios in which individuals experience a lower level of risk than that associated with many HRI tasks (Yanco et al., 2016).

Understanding trust in HRI requires that we consider the unique experience of interacting with a physically embodied, robotic agent. This effect is likely to be most profound during proximate HRI, especially in situations where a human and robot have to physically coordinate their actions to achieve respective and/or joint goals. Proximity may even heighten human experiences of vulnerability, increasing the influence of trust on performance outcomes. These "performance-based", proximate

interaction scenarios (Lewis, Sycara, & Walker, 2018) are the primary focus of this chapter. As one might expect, in performance-based interactions the kinds of performance-based robot characteristics identified by Hancock et al. (2011) have a large effect on trust (Lewis et al., 2018).

A meaningful account of trust within performance-based proximate HRI will require integration of (1) knowledge about trust from human-automation interaction, (2) awareness of the differences between human-automation interaction and HRI, and (3) acknowledgment of the most salient robot-, human-, or environment-related factors within proximate HRI. Human-automation interaction research generates meaningful information about trust that is often relevant to HRI. However, HRI tasks often involve greater risk than the human-automation interaction tasks that are frequently studied (Yanco et al., 2016). For example, human interaction with remotely controlled unmanned ground robots, like urban search and rescue, involve a very high risk of success or failure. Proximate HRI is less likely to involve this kind of risky task, but more likely to pose a physical safety risk to human teammates. For example, in the US, the Occupational Safety and Health Administration (OSHA) identified seven categories of potential hazards to humans working with robots in manufacturing contexts, such as automobile production, that can result in physical safety risks including collision, trapping, and electrical hazards (OSHA, 2006).

Robot-, human-, and environment-related factors will all contribute to people's objective and subjective experiences of their safety during proximate HRI. However, we expect that performance-based robot characteristics will still be some of the strongest predictors of trust in proximate HRI. Specifically, performance-based robot characteristics related to physical movement are most critical. The close relationship between robot movement and human safety and the apparent role of these elements within trust in proximate HRI across domains are evident in Fig. 1.

As opportunities for proximate HRI become more common, models specifically designed to account for trust in proximate HRI will support optimal teaming. In the following section, we outline a set of five proposed aims to guide the development of proximate trust-aware control. We then discuss these aims within the context of two existing trust-aware robot control frameworks.

Proposed aims for trust-aware control in proximate HRI

Based on the current understanding of trust modeling and proximate HRI, we propose a set of five aims for trust-aware control in proximate HRI.

Aim 1: Establish quantitative models that can be used online to inform robot behavior during interaction. As we discussed above, quantitative models are distinct from qualitative models in their ability to support the direct prediction of trust from robot-related, human-related, and environmental factors (Yanco et al., 2016). Quantitative models are therefore necessary for robots to properly adapt their behavior to a person's level of trust throughout an interaction. For example, if the robot estimates that the human's trust in it is too high, it could purposefully display less reliable behavior (Chen, Nikolaidis, Soh, Hsu, & Srinivasa, 2018).

Aim 2: Account for the dynamic nature of trust, allowing for changes in trust prediction based on whether the interaction is most closely aligned with a formation, dissolution, or restoration phase. Multiple studies suggest that trust often changes over the course of interaction, and that the phase of interaction may determine which factors will have the greatest influence on trust (Desai et al., 2012; Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013; Lewis et al., 2018). Therefore, to accurately predict trust, robots must weight these factors differently depending on the phase of the interaction.

Aim 3: Include the measurement of human trust and policies specifying associated changes in robot behavior based on this feedback within quantitative models that regulate robot behavior online during interaction. Different levels of trust will require varying responses from robots to maintain fluent teaming. As a result, robot behavioral policies must explicitly take trust measurements and predictions into account. Doing so will enable robots to actively support teaming, as opposed to passively observing the human's state and behaving reactively.

Aim 4: Prioritize human safety in robot control frameworks. When humans and robots are colocated, robots have the potential to physically harm people. Therefore, to ensure that people are not hurt when interacting with robots, it is imperative that the control strategies of robots involved in proximate HRI prioritize human safety. Additionally, such control strategies may make people more comfortable being physically close to the robot, which would likely lead to more fluent interactions.

Aim 5: Incorporate environmental factors along with robot- and human-related factors in models that define interactive robot behavior. Given the impact of environmental factors on trust (Hancock et al., 2011), including such factors in trust models will likely increase model effectiveness. This is especially important in dynamic environments, where changing conditions could affect trust levels regardless of human- and robot-related characteristics.

These aims allow us to assess the ability of existing trust-aware frameworks to support proximate human-robot teaming and promote the development of a new generation of trust-aware control frameworks specifically designed for proximate HRI.

Existing trust-aware control frameworks for HRI

At present, there are two existing frameworks that provide trust-aware control of robot behavior during human-robot teaming. We provide a review of these models and their ability to address our five proposed aims for trust-aware control in proximate HRI below. The relevant details of each model are also presented in Table 1.

Trust aware robot control framework (TACtiC)

Xu and Dudek (2016) presented the first trust-aware robot control framework in which real-time estimates of human trust were used to adjust robot behavior online. This framework integrated two modules: Adaptive Parameter EXploration (APEX) and the Online Probabilistic Trust Inference Model (OPTIMo). Together, these modules allow the framework to achieve robot learning of human task behavior, estimation of human trust during task execution, and changes in robot behavior following trust loss to support trust restoration. The authors evaluated the effectiveness of this strategy for human-single robot teams completing a visual navigation task in which the goal was to continuously track terrain boundaries.

The APEX module (Xu, Kalmbach, & Dudek, 2014) consisted of a vision-based boundary tracking algorithm capable of automatically adjusting system parameters based on human interventions with robot behavior. The authors did not design this module to directly assess or influence human trust during interaction, but rather to improve task performance by incorporating human variations in boundary tracking behavior within robot activity (e.g., those corresponding to differences in terrain).

The authors used the OPTIMo module (Xu & Dudek, 2015) to estimate human trust during HRI. Prior to testing the full robot control framework, the researchers customized this module to each human actor's trust tendencies through a set of initial practice boundary tracking trials. During these trials, the researchers collected data about human intervention with robot tracking, and periodic feedback from users about changes in their trust of the robot behavior (i.e., every five seconds the control interface prompted users to provide critiques about whether their trust was increasing or decreasing). Xu and Dudek (2015) used this data to establish personalized trust triggers for each individual, which they incorporated within a Dynamic Bayesian Network responsible for generating real-time trust updates when they employed the full trust-aware control framework. Ultimately, the OPTIMo inferred trust updates every 3s based on (1) the robot's current and recent task performance, (2) the human's most recent intervention state referenced to their personalized trust triggers, (3) periodic human critiques about changes in their trust toward the robot,

TABLE 1 Assessment of the two existing trust-aware robot control models with respect to use in proximate HRI.

Paper		**Xu and Dudek (2016)**	**Chen et al. (2018)**
Model		Online Probabilistic Trust Inference Model (OPTIMo) with Trust-Aware Conservative Control (TACtiC)	Partially observable Markov decision process for trust (trust-POMDP)
Robotic Platform		SL-Commander (all-terrain autonomous vehicle)	Bespoke robot arm
Collaborative Task		Terrain boundary tracking	Table-clearing
Real-Time Estimation	*Trust*	Estimates real-time unified human trust based on: - Robot performance (implicit) - Human intervening commands for robot direction and speed (implicit) - Human relative-scale critiques (explicit) - Human absolute-scale feedback (explicit)	Linear Gaussian system estimates trust based on: - Real-time robot task performance (implicit) - Real-time human intervention (implicit)
	Trust Calibration	None	Performance-maximizing policy leads to: - Robot object selection sequences that build trust (when the probability of failure is low) - Intentional robot failure (when the human exhibits high trust but the probability of robot failure is high)
Algorithmic Control	*Robot Movement*	Adaptive Parameter Exploration (APEX) uses human intervention commands to automatically adapt system parameters for boundary-tracking TACtiC adjusts robot movement in response to trust loss by: - Reducing speed - Smoothing steering signal	Trust-POMDP uses current robot performance and human intervention information to select object to move

Factors Included in Model	*Robot-Related*	- Robot boundary tracking performance - Individual human trust tendencies - Human intervention with robot behavior	- Probability of robot failure (simulation) - Human intervention with robot behavior
	Human-Related	- Human trust attitudes about robot behavior	
	Environmental	- Indirect robot adaptation to terrain change/task difficulty through robot adaptation to changes in human behavior	- Reward/failure/intervention values for object categories manually specified - Effects of object selection sequences learned from preliminary experiment

and (4) infrequent human responses to a trust survey. Thus the model incorporated the measurement of variables implicitly associated with human-robot trust as well as explicit human feedback about trust in the robot.

The TACtiC module used the trust information generated by the OPTIMo module to initiate changes in robot behavior. When OPTIMo identified a salient loss of trust, the robot behavior became more conservative in order to limit the negative effects of whatever caused the trust loss and prompt the human to engage with and assist the robot. The system exhibited conservative behavior through a reduction in movement speed and a smoothing of the steering signal. The authors evaluated the effectiveness of the TACtiC module within the three-part control framework by implementing it on an autonomous vehicle. The full trust-aware framework reliably predicted human trust, and users reported greater efficiency when interacting with a mildly conservative agent compared to one with no conservative control. The authors measured subjective efficiency using self-report items for human users' experiences of both collaboration and trust.

Trust-POMDP framework

Chen et al. (2018) presented the second trust-aware control framework, a trust-based computational model for robot decision making during human-robot teaming. In this work, the authors emphasize that robots frequently need to make decisions that would benefit from knowledge about a human's hidden mental state. Thus they characterize trust as a latent variable, which they model using a partially observable Markov decision process (POMDP). The trust-POMDP model is made up of two main modules: a human trust dynamics module and a human decision module. The authors used a data-driven approach to define each module during a table-clearing task in which a human and robot collaborated to clear five objects from a table (three plastic water bottles, one fish can, and one wine glass). To do this, they manually specified a reward function for the set of possible interactions between each of the three object categories and three potential task actions (robot clearing success, robot clearing failure, and human intervention). They then communicated this reward function to participants before collecting the model training data. The reward function was as follows: robot succeeds in clearing the bottle (1), fish can (2), or wine glass (3); robot fails in clearing the bottle (0), fish can (−4), or wine glass (−9); human intervenes with the bottle (0), fish can (0), or wine glass (0). Generally speaking, this function is based on the idea that there is a higher risk associated with robot failure when interacting with the fish can or wine glass, but also a higher reward.

In the trust dynamics module, the authors model human trust evolution over the course of HRI as a linear Gaussian system. The resulting

model relates human trust causally to robot task performance over time based on the prespecified reward function. At any given time point, the output of this module can be used within the human decision module to generate a trust-informed prediction of human behavior. This trust-based model operates on the assumption that a human's expectations of robot success will change over time, depending on their trust in the robot. The authors compared the performance of this model with a trust-free model of human behavior in which the human expectation of robot success did not change over time.

Chen et al. (2018) designed the human decision module to predict the action a human would make in response to each possible robot action based on the current trust estimate provided by the trust dynamics module. In the context of the table-clearing task, the options for human action were to allow the robot to complete an action once it was initiated or intervene in order to prevent an expected failure. When the human decision module operated based on the trust-free model, it predicted a constant likelihood of intervention for each object over the course of interaction, regardless of past robot and human behavior. In contrast, the trust-based model predicted notable changes in the probability that a person would intervene corresponding to different trust levels, especially for the highest-risk object (i.e., the wine glass). Ultimately, the human decision module used the expected human policy for each possible robot action predicted by either the trust-based or trust-free behavior model to compute an optimal robot policy. For example, the trust-based model learned that trust increases over the course of an interaction. It, therefore, calculated policies for early trials that allowed the robot to demonstrate its trustworthiness by moving less risky objects first to avoid early human interventions (e.g., moving three plastic water bottles before moving the wine glass).

The authors compared the trust-POMDP to the trust-free version of the model by implementing them both on a robotic arm. Their results revealed that human-robot team performance, measured via accumulated reward, was significantly better when the trust-POMDP controlled robot behavior. A reduction in human intervention rates drove this increase in performance. Interestingly, there was not a significant difference in the average evolution of self-reported trust between the trust-based and trust-free models. The authors determined that this was because early successes, especially for the higher risk objects, were associated with greater increases in trust than later successes.

Chen et al. (2018) acknowledge that while they only included successful robot behavior in their experimental robot test, robots are likely to fail during real-world human-robot interaction. To capture this effect, they used the learned trust dynamics and human behavior models to compute robot policy decisions for the context in which robot failure was likely.

By adjusting the reward function in this new simulation, the authors were able to induce an "information-seeking" robot policy in which the robot selected a high-risk object early on during interaction. The human response (i.e., to intervene or not) to this action indicates whether the individual has an appropriate trust calibration to the robot's failure rate. The authors also added the opportunity for the robot to fail intentionally in this simulation, allowing the robot to actively reduce over-trust. With some modifications to the trust-POMDP they were able to demonstrate that when robotic failures are possible, a reward-maximizing policy that involves strategic, trust-reducing actions can lead to better performance than a trust-maximizing policy.

Ability of existing trust-aware control frameworks to support proximate HRI

As discussed earlier, Xu and Dudek (2016) and Chen et al. (2018) presented the first trust-aware control frameworks for real-time human-robot teaming. Both frameworks are successful in demonstrating the advantage of incorporating trust-awareness within robot control. Specifically, Xu and Dudek (2016) saw a significant effect on subjective experience, including trust reports, and Chen et al. (2018) observed a substantial decrease in the rate of human intervention with robot behavior. These models share some characteristics, but are also distinct in the ways they address each of our stated aims for valuable trust-aware proximate HRI control frameworks. We summarize each of the models' contributions to these aims as follows:

Aim 1: Establish quantitative models that can be used online to inform robot behavior during interaction. Both Xu and Dudek (2016) and Chen et al. (2018) used quantitative approaches to modeling trust in HRI that incorporate modular, probabilistic models. They implemented these models within their respective robot control architectures to guide continuous robot performance and robot decision-making online during interaction.

Aim 2: Account for the dynamic nature of trust, allowing for changes in trust prediction based on whether the interaction is most closely aligned with a formation, dissolution, or restoration phase. Xu and Dudek (2016) do not explicitly include time within the framework they experimentally tested. However, they do mention wanting to incorporate updates to the user trust model, which will reflect the changes that occur as an individual gets used to working with a robot over time, or use a data-driven approach to predict long-term changes based on initial user trust thresholds. Chen et al.'s (2018) trust dynamics module includes a time series model of trust learned from preliminary data collection for the table-clearing task.

Aim 3: Include the measurement of human trust and policies specifying associated changes in robot behavior based on this feedback within quantitative models that regulate robot behavior online during interaction. Xu and Dudek (2016) use an estimate of human trust based on both robot performance and human feedback to determine when to switch to a conservative mode of robot behavior. The selection of optimal robot policies in Chen et al.'s (2018) trust-POMDP similarly reflects both robot performance and human behavior. However, the latent variable for trust in this framework is specific to the reward function for the table-clearing task and the authors note that a more multidimensional parametrization of trust that accounts for a wider range of functions and modes of automation would be advantageous. Both frameworks serve to advance trust-aware control for use in specific task contexts, but the development of a multidimensional measure that accounts for a greater proportion of trust factors will support significant advances in trust-aware control for proximate human-robot teaming. Additionally, determining which measures can be used across task contexts and which are best in a specific context will enable better control algorithms for teaming.

Aim 4: Prioritize human safety in robot control frameworks. Neither model incorporates objective or subjective measures of human safety within their estimates of trust. However, the TACtiC module (Xu & Dudek, 2016) does directly control robot movement as a function of real-time human trust. Specifically, when trust in the robot is low its movements become slower and smoother. Chen et al.'s (2018) behavioral policies specified which object to move at what time, but did not take into account the spatial orientation of the objects, human and robot arm or the kind of movement realized by the robot arm. The authors acknowledge that the movement characteristics exhibited by the robot may influence the evolution of trust.

Aim 5: Incorporate environmental factors along with robot-related and human-related factors in models that define interactive robot behavior. The APEX module of Xu and Dudek's (2016) model is capable of adapting to changes in task difficulty (i.e., terrain conditions) indirectly through adaptation to changes in human control. Chen et al. (2018) construct risk-awareness within the table-clearing task by manually specifying the rewards associated with each object and action. Their trust dynamics and human behavior modules also include information about the progression of the table-clearing task learned from preliminary data-collection.

Neither of the trust-aware HRI control frameworks we review here individually addresses all five of our proposed aims for trust-aware control in proximate HRI. However, together the frameworks provide examples of ways to approach each of these aims.

These frameworks also include elements beyond those identified by our aims that may benefit the ease and effectiveness of their use across proximate HRI contexts. For instance, Xu and Dudek's (2016) OPTIMo

module is customized to each individual using a set of training trials before the full control framework is deployed autonomously. In some ways this limits the ease of the framework's use by increasing the amount of time required to train the autonomous system for interaction with each individual user. However, the additional effectiveness afforded by the resulting individualized trust estimates is likely to make this component more of a benefit than a limitation. Additionally, Chen et al.'s (2018) trust-POMDP can actively elicit trust calibration by deliberately behaving in a way that will increase or decrease human trust. This is likely to increase the effectiveness of the framework, especially given the authors' observation that a performance-maximization policy can be superior to a trust-maximization policy, and the existing evidence that trust calibration supports effective teaming (see Ogreten et al., 2010).

The platforms and tasks used to test these control frameworks also have implications for their use across proximate HRI contexts. Xu and Dudek's (2016) autonomous driving task does involve a proximate spatial relationship between a human and robotic system. However, autonomous driving is a special case within proximate HRI. As a result, the trust dynamics between a driver and an autonomous vehicle may be different from the ones that exist during interaction with a robot arm or mobile manipulator. In the context of autonomous vehicles, the typical effects of physical proximity on trust dynamics may be more relevant to interactions with pedestrians. The fact that Xu and Dudek (2016) saw greater errors in their trust feedback estimates during the autonomous driving test compared to a previous simulation test indicates that users were more cautious during autonomous driving. We expect that being able to adapt the OPTIMo to different levels of actual and perceived risk would extend its effectiveness across a range of different proximate HRI tasks. It will also be necessary to define meaningful task-specific robot adaptation and conservative control behaviors in order to use the three-module framework for proximate HRI tasks other than boundary-tracking.

Chen et al.'s (2018) table-clearing task and robotic arm are characteristic of a number of proximate HRI tasks. However, it is important to note that both modules of the trust-POMDP include task-specific, probabilistic models that were learned from a sizeable data set. The authors mention that models for other tasks or domains can be substituted for the ones associated with the table-clearing task, but this may require additional data collection and learning which could be a time-consuming process. Relatedly, the present version of the control framework is based on a manually designated, task-specific reward function that may not actually correspond to the human conception of the task and would be irrelevant to other tasks. The authors acknowledge that a more accurate understanding of reward could be learned by the trust-POMDP through additional preliminary data collection.

At present, Chen et al.'s (2018) learned policies assume that robot capabilities will be static. In reality, it is likely that robot capabilities will vary over time or with changes in the environment, especially during proximate human-robot teaming where environments are often dynamic and noisy. Xu and Dudek (2016) have some ability to account for these changes by incorporating continuous robot learning of human control behaviors. Chen et al. (2018) also identified a large variance among users in the training data set. This indicates that the experience of trust and reliance on trust during decision-making likely varied by individual. The kind of individualization included in Xu and Dudek's (2016) framework provides an opportunity to account for these effects and improve model performance.

Discussion

Effective models of trust-aware control in proximate HRI will benefit from continued attention to the five aims outlined in the current work. The individual customization and behavioral flexibility and adaptation to human and environmental changes of Xu and Dudek's (2016) system along with the understanding of trust evolution and ability to shape trust calibration featured in Chen et al.'s (2018) work will also be especially valuable. The further advancement of trust-aware control in proximate HRI will depend on the consideration of factors that are not prioritized in the existing frameworks, especially subjective and objective human safety and robot movement behaviors as related to human safety as well as robot communication. There are some other challenges to proximate HRI that will need to be addressed to optimize trust-aware control in these contexts as well.

Advancing trust-aware control of robot movement in proximate HRI

Critically, neither of the existing trust-aware control frameworks explicitly account for human safety, which is likely to be central to the success of trust-aware, proximate interaction. It follows that the most effective trust-aware control frameworks for proximate HRI will include additional consideration of the effects of robot movement during the interaction. Attention to the trust-aware control of robot movement will not only improve the objective and subjective safety of proximate HRI, but it will also provide better opportunities for robots to improve trust calibration through clear communication of robot capabilities, robot intention, and human-robot roles and relationships.

Humans communicate a wide variety of information via natural language, but robots are often limited in their ability to process and produce communication through language. Some platforms are able to achieve human trust following system errors by communicating via text using a visual interface (Rezvani et al., 2016). However, many nonverbal behaviors that are more easily realizable by a range of robots can also convey meaningful information during human-robot teaming. For example, Lee, Knox, Baumann, Breazeal, and DeSteno (2013) demonstrated that the sequencing of nonverbal robot gestures influences human perceptions of robot trustworthiness during proximate human-robot teaming. Similarly, robot movement characteristics like speed, smoothness, and shape can communicate information about robot behavior and influence human trust during proximate human-robot teaming (Dragan, Bauman, Forlizzi, & Srinivasa, 2015; Dragan, Lee, & Srinivasa, 2013; Huang, Bhatia, Abbeel, & Dragan, 2018; Riek et al., 2010; Xu & Dudek, 2016).

Chen et al. (2018) ran a simulation in which they used intentional, movement-related task failures to reduce human trust when their trust-POMDP control framework detected over-trust in a robot. Specifically, the robot dropped an object during movement, communicating a lower level of reliability than the simulated human previously attributed to it. By simulating this policy the authors demonstrated that the effective communication of robot capabilities and appropriate calibration of human trust would ultimately support greater team performance. Other researchers proposed that similar demonstrations of robot capabilities are beneficial during preliminary training for human-robot teams. For example, researchers investigating trust in automation noted that introducing failures during training is likely to help set human expectations, which in turn supports effective trust calibration (Schaefer, Chen, Szalma, & Hancock, 2016) and reduces complacency as well as the inherent bias to trust automation (Bahner, Huper, & Manzey, 2008).

Robot movement can also be used to communicate intention. Dragan et al. (2013, 2015) demonstrated that the kind of path used to achieve goal-directed robot arm movement has a significant impact on collaborative proximate human-robot teaming. They found that *legible* movement trajectories, generated to prioritize the expression of robot intent, led to more fluent collaboration than *predictable* trajectories, which matched what a human co-actor would expect given a specific end-goal. Individuals preferred both of these movement styles to *functional* trajectories, which are based purely on reaching a goal and avoiding collisions.

Some participants leaned away from the robot more during functional movement and modified their movements to maintain greater distance (Dragan et al., 2015). This indicates that these individuals felt less safe when interacting with the robot exhibiting functional movement, although the authors note that participants reported disliking the functional movement mainly because it was difficult to coordinate their

movements within the shared task space. Regardless of whether legible movement primarily supports effortless and efficient human-robot coordination or human experiences of safety, it is likely to have a meaningful effect on trust during proximate HRI. Similarly, Che, Okamura, and Sadigh (2020) recently used a combination of robot movement characteristics and haptic feedback to communicate robot intention during a spatial navigation task. This planning framework increased users' trust in the robot compared to a simple collision avoidance algorithm.

The coordination of coactor behavior can be understood as an indicator of team cohesion. Researchers investigating HHI identified strong connections between interpersonal physical coordination and experiences of social connection such as liking and affiliation (e.g., Miles, Griffiths, Richardson, & Macrae, 2010; Miles, Lumsden, Richardson, & Neil Macrae, 2011). Such coordination also augments functional teaming dynamics through positive effects on cooperation and collaborative problem-solving (Miles, Lumsden, Flannigan, Allsop, & Marie, 2017). Launay, Dean, and Bailes (2013) showed a direct, positive association between synchronization and trust for humans coordinating their finger taps with a virtual agent. It is likely that a similar relationship exists for movement coordination between human-robot teammates. In fact, in their recent work on human-agent teaming, Wynne and Lyons (2018) identified synchrony as one of six key factors that influence human perceptions of autonomous agent teammate-likeness, the extent to which they see the agent as a capable partner rather than an instrumental tool.

Robot movement behavior can also be used to communicate information about the roles of human and robot coactors within a team structure. Within the supervisor-worker dynamic of Xu and Dudek's (2016) autonomous vehicle study, a switch to conservative robot behavior signals a need for the human to provide additional guidance. The researchers are able to elicit this guidance by having the vehicle produce a slower and smoother movement. These conservative behaviors were inspired by previous work demonstrating that humans experience robots as more teachable when they exhibit long action delays during early interaction (Tanaka, Ozeki, & Oka, 2010), as well as work indicating that humans can understand human-like hesitation gestures exhibited by robots (Moon, Parker, Croft, & Van der Loos, 2013). The ability of a robot to signal a need for additional human assistance through these kinds of movement delays and hesitation behaviors is a valuable tool for supporting appropriate trust calibration in proximate human-robot teaming.

Challenges to trust-aware control in proximate HRI

Many challenges to achieving effective and efficient HRI are especially relevant within proximate HRI. Iqbal and colleagues articulated how several of these challenges affect robot perception and action (Iqbal,

Gonzales, & Riek, 2015; Iqbal, Rack, & Riek, 2016; Iqbal, Shah, & Riek, 2018; Iqbal & Riek, 2017). First, the unpredictable nature of human behavior and human environments makes it difficult for robots to sense and understand human behavior accurately and to respond accordingly. Similarly, while robots are capable of signaling a change in their contribution to teaming (e.g., by displaying hesitation behaviors as in Xu & Dudek, 2016), they may have trouble-detecting changes in team dynamics initiated by a human coactor. This limits the ability of a human-robot team to fluently transition between role distributions during an ongoing task.

Iqbal and Riek (2017) also identify the limitations that arise as a result of limited behavioral flexibility in robots as many robots are designed to perform a single task. This is especially restrictive in proximate teaming contexts, where robots are often expected to perform multiple tasks and support trust while also approximating human social behaviors.

In addition to the technical challenges facing robot platforms for use in proximate HRI, the measurement of trust is also an obstacle to the success of trust-aware control. Recent work demonstrated that there is not always a significant relationship between trust attitudes, as measured via self-report and behavioral reliance during proximate human-robot teaming (Lohani, Stokes, McCoy, Bailey, & Rivers, 2016). This makes explicit measures of trust attitudes potentially unreliable for assessing meaningful behavioral human reliance during teaming.

The alternative is to monitor implicit measures associated with trust and reliance during ongoing human-robot teaming. Human intervention in robot behavior is a common implicit metric for reliance (e.g., Chen et al., 2018; Xu & Dudek, 2016). Researchers have also begun to identify other physiological and behavioral patterns associated with human states of trust and reliance (e.g., Khalid, Liew, Voong, & Helander's, 2018 use of facial expression, voiced speech, and heart rate).

As we discussed, it is also important to be able to capture the dynamic nature of trust over the course of proximate human-robot teaming. Yang, Unhelkar, Li, and Shah (2017) illustrated this by showing that the "trust of entirety", or trust over a person's entire experience with a robot, is better understood by examining the evolution of trust over time than by taking a single measure of trust at the end of the interaction. This evolution can be quantified by measuring the area under the curve (AUTC) for a variable, or multidimensional variable, representing trust (Desai et al., 2013). This method for estimating trust is likely to be very valuable within trust-aware control. However, it depends on identifying variables that can be measured continuously or frequently while also being accurate indicators of trust, which may be difficult depending on the proximate HRI task.

Lastly, HRI researchers only recently started to draw attention to the importance of human trustworthiness (Basu & Singhal, 2016; Takeda, 2016). For example, Takeda (2016) demonstrated that drivers who display

overreliance on an automated driving system also display greater deviation in gaze behavior. This indicates that gaze behavior may be useful for establishing the trustworthiness of a human in an automated driving context. In turn, this can inform trust-aware control decisions. The need to be able to measure human trustworthiness as well as human trust in robotic systems in order to support safe, effective control is relevant to a number of other proximate HRI contexts as well. Thus continued work on trust-aware control in proximate HRI should aim to include the measurement of both human trust in a robot coactor and human trustworthiness.

The future of proximate HRI

In this chapter, we provided the first in-depth discussion of how trust-aware control can be used to advance teaming in the kinds of proximate HRI scenarios that are becoming increasingly common across healthcare, manufacturing, and education. In doing this, we identified the trust-related factors that are most likely to be critical specifically within proximate HRI contexts. Based on these factors we outlined five primary aims for trust-aware control in proximate HRI and discussed how the elements of the two existing trust-aware control frameworks for HRI succeed or fail to address these aims.

Through this process, the importance of robot movement in trust-aware control emerged as a primary motivation for ongoing work. Robot movement characteristics can be used to communicate robot capabilities, robot intention, and human-robot roles and relationships. As we discussed, each of these movement behaviors can act to shape trust during proximate human-robot teaming. We acknowledged that there are a number of technical challenges to the fluent and flexible use of current robotic systems within proximate HRI, as well as difficulties in measuring trust in ways that are meaningful to trust-aware control frameworks. The progress of teaming within proximate HRI will require continued attention to these challenges. Ultimately, however, it is work on the trust-aware control of robot movement behavior via the methods we discussed in this chapter that will offer opportunities for transformative advances in proximate HRI.

References

Adams, B. D., Bruyn, L. E., Houde, S., & Angelopoulos, P. (2003). Trust in automated systems literature review (DRDC Toronto No. CR-2003-096). Toronto, Canada: Defence Research and Development Canada.

Bahner, J. E., Huper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001.

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008, August). The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE international symposium on robot and human interactive communication* (pp. 701–706). IEEE. https://doi.org/10.1109/ROMAN.2008.4600749.

Basu, C., & Singhal, M. (2016, March). Trust dynamics in human autonomous vehicle interaction: A review of trust models. In *2016 AAAI spring symposium series*.

Che, Y., Okamura, A. M., & Sadigh, D. (2020). *Efficient and trustworthy social navigation via explicit and implicit robot-human communication. IEEE Transactions on Robotics*. arXiv preprint arXiv:1810.11556.

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2018). Trust-aware decision making for human-robot collaboration: Model learning and planning. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction* (pp. 307–315). ACM.

De Brun, M. L., Moffitt, V. Z., Franke, J. L., Yiantsios, D., Housten, T., Hughes, A., et al. (2008). Mixed-initiative adjustable autonomy for human/unmanned system teaming. In *AUVSI unmanned systems North America conference*.

de Visser, E., & Parasuraman, R. (2011). Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231. https://doi.org/10.1177/1555343411410160.

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. A. (2013). Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on human-robot interaction* (pp. 251–258). IEEE Press.

Desai, M., Medvedev, M., Vazquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., et al. (2012). Effects of changing reliability on trust of robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction* (pp. 73–80). ACM.

Desai, M., Stubbs, K., Steinfeld, A., & Yanco, H. (2009). Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of the society for the study of artificial intelligence and the simulation of behaviour (AISB) convention, new frontiers in human-robot interaction*. https://doi.org/10.1184/R1/6552464.v1.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, *48*(3), 474–486. https://doi.org/10.1518/001872006778606822.

Dragan, A. D., Bauman, S., Forlizzi, J., & Srinivasa, S. S. (2015, March). Effects of robot motion on human-robot collaboration. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 51–58). ACM.

Dragan, A. D., Lee, K. C., & Srinivasa, S. S. (2013, March). Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE international conference on human-robot interaction* (pp. 301–308). IEEE Press. https://doi.org/10.1145/2696454.2696473.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4), 177–190.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007, May). Measurement of trust in human-robot collaboration. In *2007 International symposium on collaborative technologies and systems* (pp. 106–114). IEEE. https://doi.org/10.1109/CTS.2007.4621745.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & 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. https://doi.org/10.1177/0018720811417254.

Hayes, C. J., Moosaei, M., & Riek, L. D. (2016, August). Exploring implicit human responses to robot mistakes in a learning from demonstration task. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 246–252). IEEE. https://doi.org/10.1109/ROMAN.2016.7745138.

Huang, S. H., Bhatia, K., Abbeel, P., & Dragan, A. D. (2018, October). Establishing appropriate trust via critical states. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 3929–3936). IEEE. https://doi.org/10.1109/IROS.2018.8593649.

Iqbal, T., Gonzales, M. J., & Riek, L. D. (2015, September). Joint action perception to enable fluent human-robot teamwork. In *2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 400–406). IEEE. https://doi.org/10.1109/ROMAN.2015.7333671.

Iqbal, T., Rack, S., & Riek, L. D. (2016). Movement coordination in human–robot teams: A dynamical systems approach. *IEEE Transactions on Robotics*, *32*(4), 909–919. https://doi.org/10.1109/TRO.2016.2570240.

Iqbal, T., & Riek, L. D. (2017). Human-robot teaming: Approaches from joint action and dynamical systems. In *Humanoid robotics: A reference* (pp. 2293–2312). https://doi.org/10.1007/978-94-007-7194-9 137-1.

Iqbal, T., Shah, A., & Riek, L. D. (2018). Toward a real-time activity segmentation method for human-robot teaming. In *Proc. of the robotics: Science and systems (RSS), towards a framework for joint action: What about theory of mind workshop*.

Khalid, H., Liew, W. S., Voong, B. S., & Helander, M. (2018, August). Creativity in measuring trust in human-robot interaction using interactive dialogs. In *Congress of the international ergonomics association* (pp. 1175–1190). Cham: Springer.

Lasota, P. A., & Shah, J. A. (2015). Analyzing the effects of human-aware motion planning on close-proximity human-robot collaboration. *Human Factors*, *57*, 21–33.

Launay, J., Dean, R. T., & Bailes, F. (2013). Synchronization can influence trust following virtual interaction. *Experimental Psychology*, *60*(1), 53–63. https://doi.org/10.1027/1618-3169/a000173.

Lee, J. J., Knox, B., Baumann, J., Breazeal, C., & DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in Psychology*, *4*, 893. https://doi.org/10.3389/fpsyg.2013.00893.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/IJHC.1994.1007.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*, 50–80. https://doi.org/10.1518/hfes.46.1.50 30392.

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In *Foundations of trusted autonomy* (pp. 135–159). Springer.

Li, D., Rau, P. P., & Li, Y. (2010). A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, *2*(2), 175–186.

Lohani, M., Stokes, C., McCoy, M., Bailey, C. A., & Rivers, S. E. (2016, March). Social interaction moderates human-robot trust-reliance relationship and improves stress coping. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 471–472). IEEE. https://doi.org/10.1109/HRI.2016.7451811.

Marble, J. L., Bruemmer, D. J., Few, D. A., & Dudenhoeffer, D. D. (2004, January). Evaluation of supervisory vs. peer-peer interaction with human-robot teams. In *Proceedings of the 37th annual Hawaii international conference on system sciences, 2004.* IEEE. https://doi.org/10.1109/HICSS.2004.1265326. 9 pp.

Matsumoto, S., Washburn, A., & Riek, L. D. (2018). Human-robot co-manipulation demonstration. In: *Technology showcase for the contextual robotics institute forum*.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335.

McKendrick, R., Shaw, T., de Visser, E., Saqer, H., Kidwell, B., & Parasuraman, R. (2014). Team performance in networked supervisory control of unmanned air vehicles: Effects of automation, working memory, and communication content. *Human Factors*, *56*(3), 463–475. https://doi.org/10.1177/0018720813496269.

Miles, L. K., Griffiths, J. L., Richardson, M. J., & Macrae, C. N. (2010). Too late to coordinate: Contextual influences on behavioral synchrony. *European Journal of Social Psychology*, *40*(1), 52–60. https://doi.org/10.1002/ejsp.721.

Miles, L. K., Lumsden, J., Flannigan, N., Allsop, J. S., & Marie, D. (2017). Coordination matters: Interpersonal synchrony influences collaborative problem-solving. *Psychology*, *8*, 1857–1878.

Miles, L. K., Lumsden, J., Richardson, M. J., & Neil Macrae, C. (2011). Do birds of a feather move together? Group membership and behavioral synchrony. *Experimental Brain Research*, *211*(3), 495–503. https://doi.org/10.1007/s00221-011-2641-z.

Moon, A., Parker, C. A., Croft, E. A., & Van der Loos, H. F. (2013). Design and impact of hesitation gestures during human-robot resource conflicts. *Journal of Human-Robot Interaction*, *2*(3), 18–40. https://doi.org/10.5898/JHRI.2.3.Moon.

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control*, *21*(4–5), 203–211. https://doi.org/10.1177/014233129902100408.

Mori, M. (1970). The uncanny valley, translated by K. F. MacDorman and T. Minato. *Energy*, *7*(4), 33–35.

Muir, B. M. (1989). *Operators' trust in and percentage of time spent using the automatic controllers in supervisory process control task.* Doctoral dissertationUniversity of Toronto.

Occupational Safety and Health Administration (2006). *OSHA technical manual TED 01-00 015.* Washington, DC: US Department of Labor.

Ogreten, S., Lackey, S., & Nicholson, D. (2010, May). Recommended roles for uninhabited team members within mixed-initiative combat teams. In *2010 International symposium on collaborative technologies and systems* (pp. 531–536). IEEE. https://doi.org/10.1109/CTS.2010.5478468.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013, March). Building appropriate trust in human-robot teams. In *2013 AAAI spring symposium series*.

Paepcke, S., & Takayama, L. (2010, March). Judging a bot by its cover: An experiment on expectation setting for personal robots. In *2010 5th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 45–52). IEEE.

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55.

Rezvani, T., Driggs-Campbell, K., Sadigh, D., Sastry, S. S., Seshia, S. A., & Bajcsy, R. (2016, November). Towards trustworthy automation: User interfaces that convey internal and external awareness. In *2016 IEEE 19th International conference on intelligent transportation systems (ITSC)* (pp. 682–688). IEEE.

Riek, L. D. (2017). Healthcare robotics. *Communications of the ACM*, *60*(11), 68–78.

Riek, L. D., Rabinowitch, T. C., Bremner, P., Pipe, A. G., Fraser, M., & Robinson, P. (2010, March). Cooperative gestures: Effective signaling for humanoid robots. In *Proceedings of the 5th ACM/IEEE international conference on human-robot interaction* (pp. 61–68). IEEE Press.

Riley, V. A. (1994). Human use of automation. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *55*(6-B), 2425.

Robinette, P., Howard, A. M., & Wagner, A. R. (2015, October). Timing is key for robot trust repair. In *International conference on social robotics* (pp. 574–583). Springer.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015, March). 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* (pp. 141–148). ACM.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228.

Shaw, T., Emfield, A., Garcia, A., de Visser, E., Miller, C., Parasuraman, R. , et al. (2010). Evaluating the benefits and potential costs of automation delegation for supervisory control of multiple uavs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(19), 1498–1502. https://doi.org/10.1177/154193121005401930.

Takeda, K. (2016). Modeling and detecting excessive trust from behavior signals: Overview of research project and results. *Human-harmonized information technology* (pp. 57–75). Vol. 1 (pp. 57–75). Tokyo: Springer. https://doi.org/10.1007/978-4-431-55867-5 3.

Tanaka, K., Ozeki, M., & Oka, N. (2010, March). The hesitation of a robot: A delay in its motion increases learning efficiency and impresses humans as teachable. In *Proceedings of the 5th ACM/IEEE international conference on human-robot interaction* (pp. 189–190). IEEE Press.

Washburn, A., Adeleye, A., An, T., & Riek, L. D. (2020). Robot errors in proximate HRI: How functionality framing affects perceived reliability and trust. *ACM Transactions on Human-Robot Interaction (THRI)*, *9*(3), 19.

Woodworth, B., Ferrari, F., Zosa, T. E., & Riek, L. D. (2018). Preference learning in assistive robotics: Observational repeated inverse reinforcement learning. F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), *Proceedings of the 3rd machine learning for healthcare conference* (pp. 420–439). *Proceedings of machine learning research: Vol. 85*(pp. 420–439). Palo Alto, CA: PMLR.

Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, *19*(3), 353–374.

Xu, A., & Dudek, G. (2015, March). Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 221–228). ACM. https://doi.org/10.1145/2696454.2696492.

Xu, A., & Dudek, G. (2016, October). Maintaining efficient collaboration with trust-seeking robots. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 3312–3319). IEEE. https://doi.org/10.1109/IROS.2016.7759510.

Xu, A., Kalmbach, A., & Dudek, G. (2014, May). Adaptive parameter EXploration (APEX): Adaptation of robot autonomy from human participation. In *2014 IEEE international conference on robotics and automation (ICRA)* (pp. 3315–3322). IEEE. https://doi.org/10.1109/ICRA.2014.6907336.

Yanco, H. A., Desai, M., Drury, J. L., & Steinfeld, A. (2016). Methods for developing trust models for intelligent systems. In *Robust intelligence and trust in autonomous systems* (pp. 219–254). Springer.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017, March). Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 408–416). IEEE. https://doi.org/10.1145/2909824.3020230.