10 Human–Robot Interaction

Connor Esterwood, Qiaoning Zhang, X. Jessie Yang, and Lionel P. Robert

10.1 INTRODUCTION

Organizations across our society are increasingly relying on robots to engage in interactions with humans. A robot can be defined as a sophisticated machine that is equipped with sensors, processing capabilities, and actuators that enable it to perceive, analyze, and interact with its surroundings in a physical manner (You et al., 2018; You and Robert Jr, 2018). Human–robot interaction (HRI) is an area of research that focuses on identifying and understanding the factors that promote or hinder human interaction with robots. The study of HRI is multidisciplinary and involves fields such as psychology, information science, computer science, engineering, and design and has the potential to transform various fields of human endeavor such as finance, manufacturing, health care, and education. At its core, HRI research seeks to design robots that are more responsive, engaging, and trustworthy to promote their acceptance by humans. This includes conducting research that identifies ways to build robots that are capable of interacting with humans in an intuitive, comfortable, and natural way to help promote collaborations between humans and robots.

The goal of this chapter is to provide a comprehensive overview of the most vital areas shaping HRI. To accomplish this, this chapter is organized in the following way. First, it presents and discusses the types of robots. The field of HRI has explored a diverse range of robots, revealing their impact on various outcomes (Robert, 2018; Robert Jr et al., 2020). Second, this chapter presents a scoping literature review that surveys trust in HRI. Trust is the foundation by which humans have sought, engaged, and benefited from one another. It is no surprise that trust has been shown to be vital for collaborative action between humans and robots. Third, this chapter identifies and discusses the literature on personality in HRI. Personality, both human and robot, has been shown to impact the interactions between humans and robots. Personality can be viewed as a representation of an individual human or robot's future behaviors, cognitions, and emotional reactions (Robert, 2018). To fully grasp the intricacies of personality in HRI, this chapter employs a multidisciplinary approach encompassing several views on both human and robot personality. Next, this chapter presents the literature on robot explanations. "Robot explanations" can be defined as the reasons that the robot provides to make its actions clear or easy to understand (Zhang et al., 2021). Robot explanations can decrease the uncertainty associated with the robot's actions by providing transparency. Finally, this chapter delves into the various metrics used to evaluate HRI. We also discuss the latest research on evaluation metrics of human interaction with robots. Traditionally, evaluation metrics of human interaction with robots have survey-based static measures. However, recent advances in sensors allow for real-time measures based on physiological changes, which can be obtained alongside or in place of traditional survey measures. In summary, this chapter provides an overview of important HRI areas shaping the field today.

10.2 ROBOTS USED IN HRI RESEARCH

10.2.1 Type of Robot

There are many definitions of the term "robot." The definition that best aligns with the use of robots in the HRI field is offered by You and Robert Jr (2018), who defined robots as technologies that can have either virtual or physical-embodied actions. As You and Robert Jr (2018) pointed

DOI: 10.1201/9781003490685-10

out, embodiment and representation of embodied behaviors are what make robots different from other artificial intelligence (AI) technologies. Consistent with this definition, the field of HRI has utilized a wide range of robotic platforms, including custom one-off designs and standardized commercially available robots, with the latter being especially important for ensuring replication and reproducibility. Generally, at least 20 types of robots have been employed in HRI studies (Robert, 2018; Zimmerman et al., 2022), highlighting the diverse array of robotic platforms present in the literature. One way to categorize these robots is by their morphology, which can be classified as either humanoid or non-humanoid. Research has shown that differences between these morphologies can have a significant impact on HRI outcomes (Robert, 2018; Robert Jr et al., 2020). In the following sections, we dive deeper into these two types of robots and provide a brief overview of the most common robots used to represent each type.

10.2.1.1 Humanoid Robot

Humanoid robots, engineered to mimic human form and behavior, offer crucial insights for HRI research. These robots are defined by their human-like appearance, typically featuring a head, two arms, two legs, and a torso (Hirai et al., 1998; Ishida et al., 2001). Their configuration enables them to execute tasks that closely resemble human actions, although some humanoid robots only replicate specific body parts to concentrate on certain aspects of human-like interaction.

The expansive category of humanoid robots encompasses two primary subcategories: avatars and human-like robots. Avatars can manifest as virtual representations or physical platforms, conveying responses based on simulated facial expressions, gaze, or other cues (Zimmerman et al., 2022). These embodiments simulate human presence and interaction, enabling researchers to investigate human perceptions and responses to humanoid representations in various contexts. On the other hand, human-like robots are physical machines explicitly engineered to resemble humans in both form and behavior. Figure 10.1 shows that among the humanoid robots used in research, the Nao robot, Pepper robot, and Baxter robot have been identified as the most popular choices across multiple studies (Robert, 2018; Zimmerman et al., 2022).

10.2.1.2 Non-Humanoid Robot

Non-humanoid robots typically have simpler embodiments that are targeted toward specific tasks or domains (Cha et al., 2018; Coeckelbergh, 2011; Terada et al., 2007). As a result, these robots are utilized in a wide variety of settings and applications, such as health care, agriculture, space,

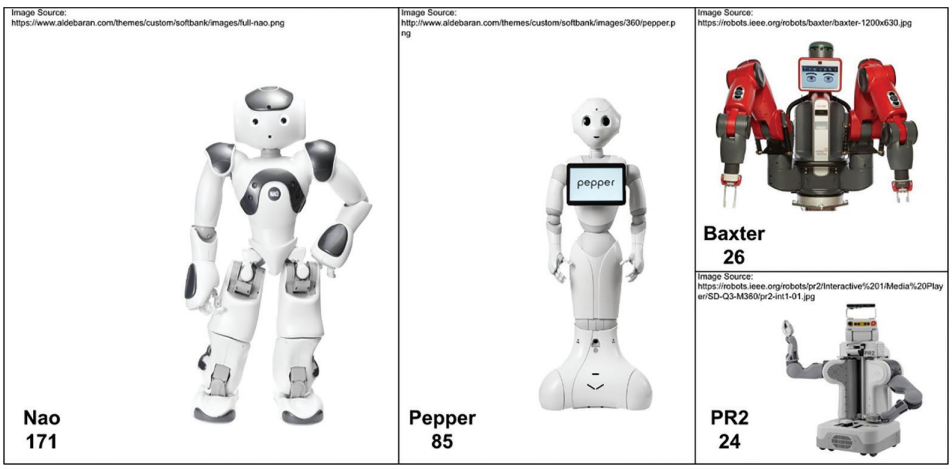

FIGURE 10.1 Figure illustrates the most popular humanoid robots identified by Zimmerman et al. (2022) and the number of times they were used across the human–robot interaction (HRI) literature between 2015 and 2021.

industry, automotive, and service (Cha et al., 2018). These robots encompass a variety of forms, shapes, and functionalities, with numerous robot types falling under this classification. As identified by Zimmerman et al. (2022), there are 15 distinct types of robots classified as non-humanoid robots. Among these are aquatic robots, which operate in aquatic environments (Georgiades et al., 2004; Long et al., 2006), and audio-only robots that interact with humans primarily through auditory means using voice commands and responses, without a physical or visual presence (Guinness et al., 2019; Raghunath et al., 2021). Other non-humanoid robot types include drones, which are capable of flying and performing tasks like aerial photography, surveillance, and delivery (Bhat et al., 2022; Suzuki, 2018); robotic hands, which are specialized robots that replicate human hand functionality (Piazza et al., 2019); non-industrial arms, which can be used in non-industrial settings (Millo et al., 2021); and image/video robots, which utilize visual information for communication or interaction (Raghunath et al., 2021). Industrial manipulators, also known as robotic arms, are employed in manufacturing and production settings for tasks (Zanchettin et al., 2013). Industrial mobile robots are autonomous machines designed to navigate complex industrial environments (Schneier et al., 2015). Mobile manipulators combine mobility and manipulation capabilities, enabling them to navigate and interact effectively within their environment (Bostelman et al., 2016). Mobile platforms, which move on wheels or tracks, navigate various environments for tasks such as transportation, exploration, or assistance (Sørensen et al., 2015). Mobility assistant robots are specifically designed to help individuals with mobility challenges, such as wheelchair users or those with physical impairments, by providing support and guidance (Geravand et al., 2016). Simulation/video game robots exist within virtual environments and serve as interactive characters or elements (Roitberg et al., 2021). Telepresence robots facilitate remote users' virtual presence in a different location, enabling interaction (Tsui et al., 2011). Toy robots, designed for entertainment and education, have various shapes, sizes, and functionalities (Michaud et al., 2000). Last, written vignettes are textual descriptions of robot behavior or interactions used in studies where the actual robot is not present or necessary for the experiment (Moyle et al., 2013).

In the HRI field, research focusing on non-humanoid robots frequently showcases prominent robot types such as mobile manipulators, virtual or gaming robots, and industrial manipulators. As illustrated in Figure 10.2, the UR5, Sawyer, and UR10 robots have been recognized as the leading choices across various HRI studies.

10.2.2 Physically Present V.S. Virtually Represented Robots

Within the HRI field, the traditional approach of presenting physical/real-world robots to subjects in laboratory and naturalistic settings is increasingly being challenged by the use of virtual representations of robots. These virtual representations include two-dimensional and three-dimensional videos, interactive game-based environments, and virtual reality, and have gained popularity as a result of recent advances in simulation and increased accessibility of game engines and computer graphics (Mara et al., 2021). Although the coronavirus disease 2019 (COVID-19) pandemic accelerated the adoption of virtual representations of robots, researchers have been exploring this alter-native method for various reasons, including cost, complexity, unpredictability, and difficulties in programming physical robots (Esterwood et al., 2023). Virtual representations provide greater methodological flexibility, allowing for more extensive opportunities for manipulating robot characteristics and behaviors, which can broaden the range of research questions that can be explored.

Studies have found minimal differences in humans' overall experience with robots between physical/real-world robots and the same robots presented in a virtual format. However, differences have been observed in humans' perceptions of a robot's utility, immediacy, perceptions and attitudes, and performance (Kamide et al., 2014; Liang and Nejat, 2022; Mara et al., 2021). The suitability of virtual representations of robots as proxies for physical/real-world robots in HRI is under investigation. Moderating factors, such as the type of robot and the context of the study, may play a role in determining the parity between the two formats (Liang and Nejat, 2022).

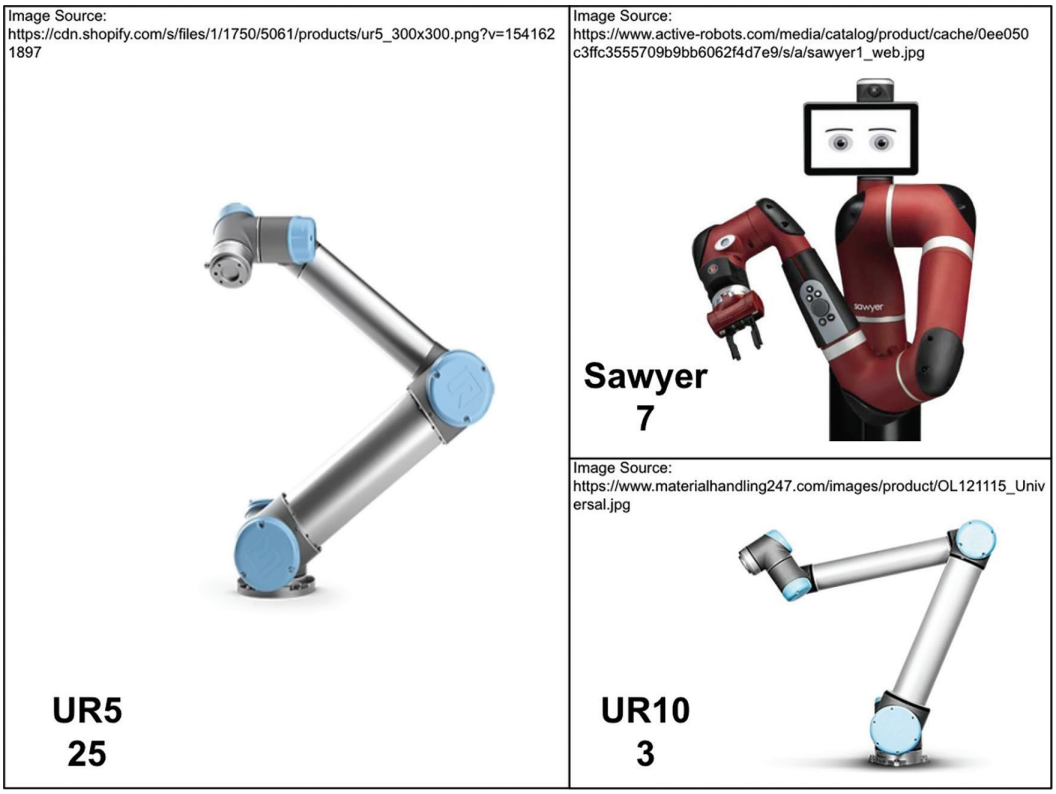

FIGURE 10.2 Figure illustrates the most popular non-humanoid robots identified by Zimmerman et al. (2022) and the number of times they were used across the human–robot interaction (HRI) literature between 2015 and 2021.

While some integrative work has been done in this space, such as Liang and Nejat (2022)'s investigation, the scope of this work has been limited to assistive robots in health care and well-being settings. Further research is needed to provide stronger support for or against the parity of physical/real-world robots and virtual representations of robots in the broader context of HRI research. Unfortunately, the current literature does not provide enough data for lower-level meta-analyses, making it challenging to draw firm conclusions. Therefore, it is crucial for researchers to continue exploring the potential benefits and limitations of both methods to better understand the use of virtual representations of robots in HRI research.

10.3 TRUST IN HRI

Trust is a vital component of any effective HRI. This is because without trust, humans fail to fully leverage robots. For example, if humans do not trust a robot, they are less likely to rely on that robot to perform the sorts of tasks that make robots useful. As a result, the benefits of robots are drastically reduced. This can lead to scenarios where work arrangements become not only unproductive but ultimately damaging to the overall productivity and well-being of workers. This is especially true when one considers recent shifts in the role of robots in working arrangements. Specifically, the roles that robots play in workplaces are shifting from tools and to teammates (You and Robert 2018), and, as a result, the various psychological and social aspects that lead to effective human–human teams are increasingly present in heterogeneous human–robot teams. Indeed, the importance of trust in this regard has not gone unnoticed, and a wealth of literature on this topic has begun

to emerge from HRI literature. In response, this section introduces the concept of human–robot trust and provides a summary of the factors that impact when robots are seen as worthy of trust. In addition, we discuss recent computational work for estimating human–robot trust in real time. We close this section with an introduction to the emerging field of human–robot trust management and recovery. In doing so, we provide a starting point for those seeking to learn more on these topics.

10.3.1 Human–Robot Trust and Trustworthiness

Trust is a complex and multifaceted aspect of HRI. As a result, no single universally accepted definition for trust—either in general or in the context of HRI—has been established. Three common definitions of trust, however, have been gaining popularity across HRI literature. These definitions stem from different fields but overlap in several important places while diverging in others.

The oldest of these definitions is that of Mayer et al. (1995). They defined trust as "The **willingness** of a party to be **vulnerable** to the actions of another party based on the **expectation** that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, Pg.712), emphasis ours. This definition stems from Human–Human Interaction (HHI) literature and contains three distinct elements. First is the central concept of vulnerability or risk, second is the positioning of trust as pre-behavioral, and third is the role of expectations. The first of these (i.e., vulnerability) is a vital component of trust because vulnerability implies risk (Robert et al., 2009). Indeed, without risk or the potential for "something of importance to be lost," trust is ultimately unnecessary, unneeded, and relatively meaningless (Mayer et al., 1995).

The vulnerability element of trust is the most widely included in subsequent trust definitions, with the remaining two popular definitions of trust in HRI implying or explicitly including vulnerability. Specifically, Lee and See (2004) referenced vulnerability in their definition of trust, which is *"The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and **vulnerability**"*. Hancock et al. (2011), on the other hand, implies vulnerability in their trust definition by defining trust as: *"The reliance by an agent that actions **prejudicial to their well-being** will not be undertaken by influential others"* (emphasis ours in both statements). In both cases, the risk is ultimately incurred by a trustor in the form of the trustee's potential actions. This, therefore, makes the trustor vulnerable to the trustee. The second element of trust based on Mayer et al. (1995)'s definition is the distinction between trusting behaviors and trust itself. In their words, with our emphasis: *"The fundamental difference between trust and trusting behaviors is between a **willingness** to assume risk and actually **assuming** risk."* (Mayer et al., 1995, Pg.724). This distinction allows one to draw a line between trust and trust-related outcomes Robert et al. (2009). This is important because not all risk-taking behaviors are trust- dependent. Other definitions of trust in HRI have been less consistent on this point. For example, although Lee and See (2004)'s definition of trust places trust as an attitude that is conceptually closer to willingness and distinct from behavior, Hancock et al. (2011)'s definition of trust as "reliance" could be interesting as a form of risk- taking or behavior. This is not to say that Hancock et al. (2011)'s definition of trust is incorrect, however, but rather that it may be inconsistent with Mayer et al. (1995) and Lee and See (2004)'s definitions when examined at a deeper level.

Finally, the third element of trust based on Mayer et al. (1995)'s definition relates to expectations. These expectations are synonymous with the concept of trustworthiness. Trustworthiness can be defined as "a multifaceted construct that captures the competence and character of the trustee" (Colquitt et al., 2007, Pg.909). Trustworthiness is distinct from trust and largely precedes it (Mayer et al., 1995). It does so by influencing the expectations that one has of a trustee and therefore their willingness to trust. For example, a trustor is more disposed to trusting a trustee who seems trustworthy but not a trustee who appears untrustworthy. What makes someone or something trust- worthy, however, is more complex, but research in HHI has provided some useful frameworks.

Generally, the most popular framework for trustworthiness divides it into three components. These are ability, integrity, and benevolence. Ability is the skillfulness or competency that trustees are believed to have at their disposal (Mayer et al., 1995). In HRI, this is a human's belief that a robot can do what it has promised. Integrity is the degree to which the trustee is seen as honest and adherent to an acceptable set of principles (Kim et al., 2020, Pg.2). In HRI, this is a human's belief that a robot is honest and acts in a morally consistent manner. Finally, benevolence is "the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive" (Mayer et al., 1995, Pg.718). In HRI this is a human's belief that a robot is acting for the human's benefit and is free from conflicts of interest. In general, the lower these expectations are, the less disposed the trustor is to be vulnerable to and rely on (i.e., trust) the trustee (Colquitt et al., 2007). This has been found to be true as much for humans (Colquitt and Salam, 2012; Colquitt et al., 2007; Poon, 2013) as for robots (Esterwood and Robert, 2021). Many factors can influence trustworthiness and by extension trust, and a wealth of research in HRI exists on this topic.

Ultimately, a range of definitions and conceptualizations exist on trust. Each of these approaches differs and ongoing debate among these and other definitions persists. These debates are increasingly common and have begun to emerge not only in the field of HRI but across multiple domains. In this chapter, however, we focus on the three increasingly popular definitions of trust in HRI that we highlighted in italics. With these established, however, another question comes to mind. Specifically, what factors influence trust in robots? Fortunately, a great deal of research has been conducted and a series of meta-analyses hold special insight that may help us answer this question.

10.3.2 Factors Influencing Trust in HRI

Across the HRI literature, an ever-increasing number of studies have sought to determine which factors in HRI can influence human–robot trust. Generally, these studies can be grouped into three distinct categories (Hancock et al., 2011; Sanders et al., 2011). These are human-related factors, robot-related factors, and contextual-related factors (Hancock et al., 2011). A summary of these factors is presented in Figure 10.3 based on the conceptual model proposed by Sanders et al. (2011).

This model has been empirically examined via tow sequential meta-analysis (Hancock et al., 2011), and the analysis provided support for a handful of the factors proposed by Sanders et al. (2011). With regard to human-related factors, these factors appeared to significantly impact trust overall, but, on closer examination only one appeared significant (Hancock et al., 2021). In particular, only factors associated with a human's characteristics as opposed to abilities appeared to significantly impact trust. Furthermore, within this sub-factor, only satisfaction, expectancy, comfort, and personality appeared to be significantly influential (Hancock et al., 2021).

For robot-related factors, these factors can be subdivided into performance-based and attribute-based factors. Overall, both of these sub-factors appear to be significantly influential in combination and when examined individually (Hancock et al., 2021). Within these sub-factors, however, the impact is not equally distributed, nor is it always positive. In particular, only the performance-based factors of dependability and reliability significantly influence trust, with dependability actually having a negative impact (Hancock et al., 2021). Furthermore, for attribute-based factors, only robot personality significantly influences trust (Hancock et al., 2021).

Finally, for contextual factors, these factors can be subdivided into collaboration- based and tasking-based sub-factors. Overall, these factors do not appear to significantly influence trust in robots. When examined individually, however, collaboration- based sub-factors do (Hancock et al., 2021). Across Hancock et al. (2021)'s meta- analysis, however, relatively few studies in this category made firm conclusions on the true impact of such factors.

Taken together, Hancock et al. (2021)'s meta-analysis and review of the antecedents of trust in HRI point to a handful of significant factors that can influence trust. In particular, it appears that a human's comfort, expectancy, personality traits, and satisfaction are influential alongside a robot's personality, dependability, and reliability. Furthermore, contextual factors such as collaboration

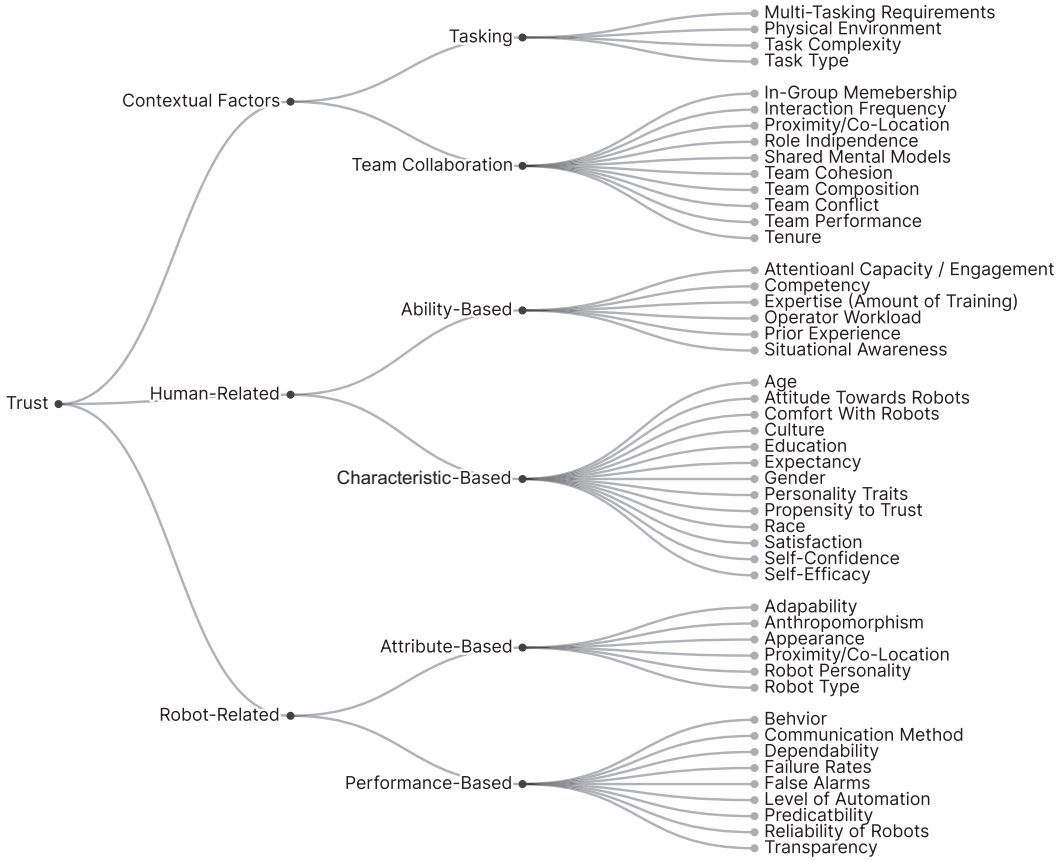

FIGURE 10.3 Hancock et al.'s 2021 model of human–robot trust and the factors that influence it.

may also be important. While these results highlight important trust-relevant factors for human–robot trust, they are by no means the only factors that may be useful. In addition, interactions among factors may exist, making some factors only relevant considering others. As a result, more research is needed, but the factors highlighted in Figure 10.4 may be of use to researchers and designers alike when considering this future work.

10.3.3 Trust Dynamics and Computational Trust Models in HRI

As described in Section 10.2.2, a growing body of research is identifying factors influencing one's trust in automation. The majority of this body of research adopts a snapshot view of trust and evaluates a person's trust at specific points in time, usually at the end of an experiment. This snapshot view, however, does not acknowledge that trust can change as a result of continual interactions with autonomy. As shown in Figure 10.5, at time *t*, agents *A*, *B*, and *C* have the same level of trust. However, their trust dynamics are different if examined over the time horizon.

Therefore, more recently another line of research has emerged that focuses on understanding the dynamics of trust formation and evolution when a person interacts with autonomy repeatedly (de Visser et al., 2020; Guo and Yang, 2021; Yang et al., 2021). Empirical studies have investigated how trust strengthens or decays as a result of moment-to-moment interactions with autonomy (Lee and Moray, 1992; Manzey et al., 2012; Moray et al., 2000; Yang et al., 2021). Based on these studies, Yang et al. (2023) summarized three major properties of trust dynamics: *continuity*, *negativity bias*, and *stabilization*. Continuity means that trust at the present moment is significantly associated with

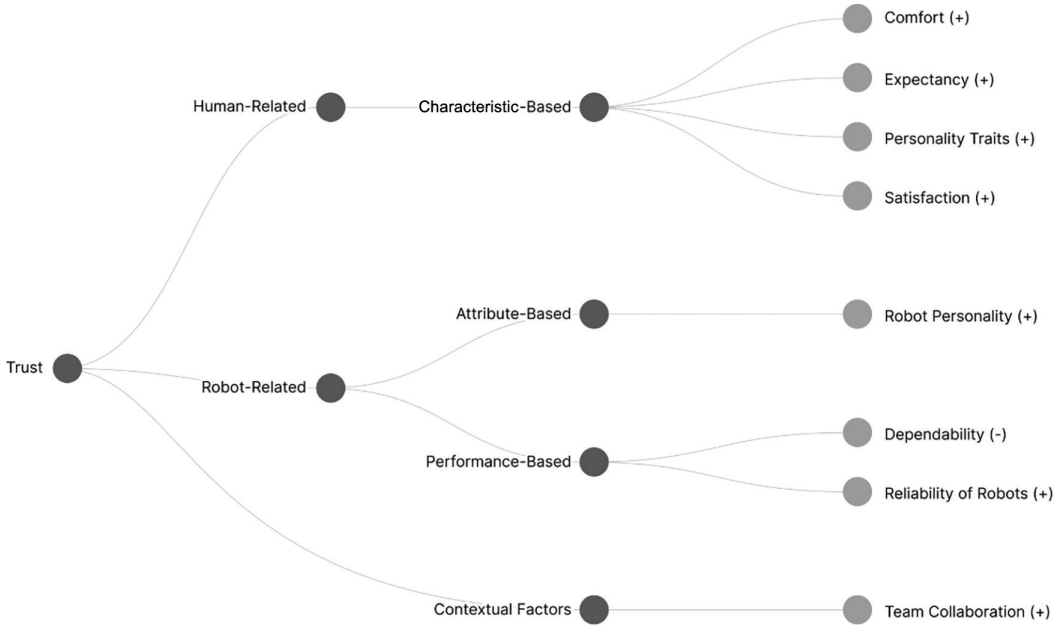

FIGURE 10.4 Hancock et al.'s 2021 model of human–robot trust and the factors that influence it where only significant effects are included. (+) indicates a positive impact on trust, (–) indicates a negative impact on trust.

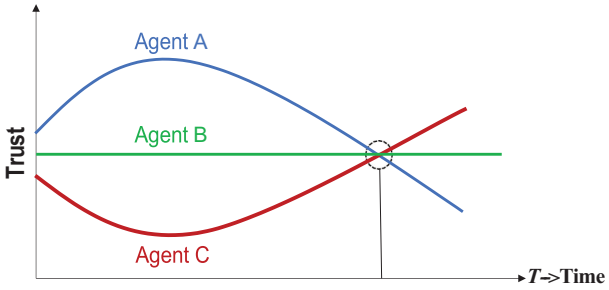

FIGURE 10.5 The static "snapshot" view versus the dynamic view of trust. At time *t*, Agents *A*, *B*, and *C* have the same level of trust. However, their trust dynamics are different.

trust at the previous moment and increases upon automation successes and decreases upon automation failures. Negativity bias means that negative experiences related to autonomy failures have a greater influence on trust than positive experiences related to autonomy successes. Stabilization means that a person's trust stabilizes over repeated interactions with the same autonomy.

Acknowledging that trust is a dynamic variable, several computational models of trust in automation have been developed. Lee and Moray (1992) proposed an auto- regressive moving average vector (ARMAV) time series model of trust that calculated trust at the present moment t as a function of trust at the previous moment $t-1$, task performance, and the occurrence of automation failures. Yang et al. (2017) examined how trust in automation evolved as an average human agent gained experience interacting with robotic agents. Their results showed that the average human agent's trust in automation stabilized over repeated interactions, and this process can be modeled using a first-order linear time-invariant dynamic system (Yang et al., 2017). Hu et al. (2016) proposed to predict trust as a dichotomy of trust vs. distrust by analyzing the human agent's electroencephalography (EEG) and galvanic skin response (GSR) data. Similarly, Lu and Sarter (2019) proposed the

use of eye-tracking metrics, including fixation duration and scan path length, to infer a human's real-time trust. Their follow-up study (Lu, 2020) used three machine-learning techniques, logistic regression, *k*-nearest neighbors (kNN), and random forest to classify the human's real-time trust level.

Instead of using physiological signals, Xu and Dudek (2015) developed the online probabilistic trust inference model (OPTIMo) utilizing Bayesian networks to estimate human trust from automation's performance and human behaviors. Building from the three empirical properties of trust dynamics (i.e., continuity, negativity bias, and stabilization). Guo and Yang (2021) proposed to model trust as a beta random variable and predict trust value in a Bayesian framework. Guo and Yang (2021) compared their model prediction results against the two models developed by Lee and Moray (1992) and Xu and Dudek (2015) and showed that their model significantly outperformed the other two models. Moreover, given that the model complies with the empirically identified properties, it guarantees good explainability and generalizability. Using Gaussian processes, Soh et al. (2020) proposed a multi-task trust transfer model that can predict human trust in a robot's capabilities across multiple tasks. More recently, Guo et al. (2023a, b) proposed the trust inference and propagation (TIP) model, the first mathematical framework for computational trust modeling in multi-human multi-robot teams. The authors asserted that there are two types of experiences that any human agent has with any robot in a multi-human multi-robot team: direct and indirect experiences. The TIP model explicitly accounts for both types of experiences and successfully captures the underlying trust dynamics, significantly outperforming a baseline model (Guo et al., 2023a, b).

The ability of a robot to accurately estimate a human's trust level in real time has led to the research of trust-ware HRI, wherein a robot can adapt its behavior in accordance with a human's trust (Azevedo-Sa et al., 2021a). Similarly, Xu and Dudek (2016) proposed a framework for using an estimated trust for trust-aware conservative control (TACtiC), in which an autonomous agent can momentarily change its behavior to be more conservative if the human's trust falls too low. In Akash et al. (2019), a trust-workload partially observable Markov decision process (POMDP) model was trained and solved to generate optimal policies for a robot to control its transparency to improve the performance of the human–robot team. Further, Chen et al. (2020) presented a trust-POMDP model that can be solved to generate optimal policies for the robot to calibrate the human's trust and improve team performance. Guo et al. (2021) presented a reverse psychology model of human trust behavior and compared it with the disuse model.

The abovementioned work focuses on the human's trust in the autonomous/robotic agent. Recently, the first bi-directional trust model that encompasses both the human's trust in the robotics agent and the robotic agent's trust in the human was developed by Azevedo-Sa et al. (2021b). In the work, Azevedo-Sa et al. (2021b) introduced a novel capabilities-based bi-directional multi-task trust model that can be used for trust prediction from either a human or a robotic trustor agent. This model is useful for control authority allocation applications that involve both the human's trust in the robot and the robot's trust in the human.

10.3.4 Trust Management and Recovery in HRI

As mentioned, many factors influence trust in robots. Trust, however, is not fixed and changes over time (de Visser et al., 2020; Guo and Yang, 2021; Yang et al., 2023). Trust increases when robots meet expectations and decreases when they do not. Although increases in trust are relatively easy to manage, decreases can have lasting effects and are hard to recover from (Lewicki and Brinsfield, 2017). Generally, such decreases are the result of violations of trust. Trust violations can take multiple forms but generally result from a trustee failing to meet the expectations of a trustor (Esterwood and Robert, 2022b). This produces a reduction in the trustor's willingness to be vulnerable and therefore trust (Costa et al., 2018; Esterwood and Robert, 2022b; Esterwood and Robert, 2023a,b; Gillespie et al., 2021; Kim et al., 2004).

Trust violations in HRI can take on three distinct forms: ability-based, integrity- based, and benevolence-based (Grover et al., 2014). An ability-based trust violation occurs when a robot fails to meet a human's performance expectations or perform a task as assigned (Sebo et al., 2019). An integrity-based trust violation occurs when the robot breaches a human's expectation of honesty and ethical conduct (Sebo et al., 2019). Lastly, a benevolence-based trust violation arises when the robot is perceived as lacking care and fails to fulfill a human's expectation of its purpose (Esterwood and Robert, 2022b). Each of these types of violations can seriously undermine trust and result in a variety of negative outcomes.

Fortunately, trustees can use several strategies to mitigate trust decreases and restore trust. The most common of these are short-term verbal trust repair strategies, including apologies, promises, explanations, and denials (Esterwood and Robert, 2022b; Lewicki and Brinsfield, 2017). Apologies, expressions of regret or remorse for a perceived transgression, are the most widely used trust repair strategy across the literature (Esterwood and Robert, 2022b). They are believed to repair trust through encouraging forgiveness (Esterwood and Robert, 2023c). Promises, on the other hand, are statements of commitment to positive future performance (Schweitzer et al., 2006). They aim to restore trust by shifting the focus from past to future behaviors and are believed to work through encouraging forgetfulness (Esterwood and Robert, 2023).

Explanations, which are discussed in more detail in Section 10.4 of this chapter, are statements that provide clear reasoning behind a trust violation. They seek to establish a shared understanding between a trustor and trustee by conveying transparency (Esterwood and Robert, 2023; Ezenyilimba et al., 2022; Lewicki and Brinsfield, 2017; Lewicki et al., 2016; Rawlins, 2008). Explanations are hypothesized to repair trust through informing the trustor (Esterwood and Robert, 2023). Denials, on the other hand, are trust repair strategies that redirect blame or reject culpability for a trust violation (Baker et al., 2018). They aim to establish the complete innocence of the trustee by shifting blame away from them and onto another entity (Lewicki and Brinsfield, 2017). Denials are hypothesized to work through misinforming because they rely largely on inaccurate information provided by the trustee to the trustor (Esterwood and Robert, 2023).

Generally, the efficacy of these repair strategies in restoring human–robot trust is mixed, with some of these strategies appearing to be effective at certain times but not others (Esterwood and Robert, 2022b). This may largely relate to other factors acting as moderators. For example, timing (Kox et al., 2021; Robinette et al., 2015), violation type (Sebo et al., 2019; Zhang, 2021), anthropomorphism (Esterwood and Robert, 2021), attitude (Esterwood and Robert, 2022a) and severity (Correia et al., 2018) have each been shown to impact the efficacy of different trust repairs. The field of trust repair in HRI, however, is still young, and much remains unknown. In particular, additional moderators or previously unexamined main effects may more clearly explain these diverging results. Alternatively, human-related factors and individual differences might largely determine the efficacy of these repairs. Indeed, much work is ongoing, and the field of trust repair continues to expand.

10.4 PERSONALITY IN HRI

Human and robot personalities play a significant role in shaping how humans interact with and use robots (Esterwood et al., 2021b; Hancock et al., 2021; Robert Jr et al., 2020). These personalities have been found to impact not just trust, but also a wide range of other outcomes as well (Alarcon et al., 2021; Esterwood et al., 2021a; Robert Jr et al., 2020). Personality is therefore an important factor to consider in the context of HRI. In this section, we delve into the concept of personality in HRI and its impact on HRI. We start by defining personality and introducing some of the various ways in which it has been conceptualized both within the HRI literature and more generally across the personality psychology literature. We then summarize what has been found across the literature and how human personality, robot personality, and the match or mismatch between the two have impacted HRI.

10.4.1 Personality and Personality Traits

To begin this section, we must first define what personality is and briefly introduce the different ways that personality has been approached in the fields of personality psychology and HRI alike. Personality can be defined as an individual's "characteristic pattern of behavior in the broad sense (including thoughts, feelings, and motivation)" (Baumert et al., 2017, Pg.527). Generally, across the field of personality psychology, five distinct approaches to personality have garnered the most attention, and debate among these schools of thought is abundant (see McMartin (2016) for a full review). The most common approach to personality across HRI literature, however, is the trait-based approach to personality (Esterwood et al., 2021b).

The trait-based approach to personality posits that traits act as the foundational elements by which personalities are constructed (McMartin, 2016). Traits are "organized dispositions within the individual" (McMartin, 2016, Pg.30). These dispositions are seen as relatively stable and, as a result, they can be used to predict and explain various aspects of human behavior (Allport, 1937; McCrae and Costa Jr, 2008). The exact makeup and number of these traits in humans is a topic of lively debate within the personality psychology literature, and multiple sets of traits have garnered empirical support (McMartin, 2016). One commonality across each of these approaches, however, is that they see personality as multidimensional and as capable of being subdivided into specific operational variables. This has historically allowed researchers to precisely link specific traits to particular outcomes, which may partially explain the popularity of this approach (Esterwood et al., 2021b; Haslam, 2007; McMartin, 2016; Tasa et al., 2011).

The most common and widely supported approach to personality traits across many fields (Li et al., 2014; Robert, 2018; Robert Jr et al., 2020), including HRI (Esterwood and Robert, 2020; Lee and Nass, 2003; Pocius, 1991; Robert, 2018; Robert Jr et al., 2020; Völkel et al., 2019) is that of the Big Five. The Big Five personality traits are extraversion, neuroticism, openness, agreeableness, and conscientiousness (Goldberg, 1992; John et al., 2008; McCrae and Costa Jr, 2008). Extraversion is often conceived of as a spectrum with two poles, one being extraversion and the other introversion. Extraversion is the extent to which someone is outgoing, assertive, vocal, and sociable (Rhee et al., 2013), whereas introversion is the level to which an individual is timid, prefers quietness, and enjoys solitude (Driskell et al., 2006). Neuroticism is the degree to which someone is easily angered, not well-adjusted, insecure, or lacks self- confidence (Driskell et al., 2006). Openness to experience is typically described as the amount to which one is imaginative, curious, and open-minded (McCrae and Costa Jr, 1997). Agreeableness can be characterized by how cooperative and friendly someone is (Peeters et al., 2006). Finally, conscientiousness is the degree to which individuals are thorough, deliberate, and mindful of their actions (Tasa et al., 2011).

Each of these personality traits in sum comprises an individual's personality. This personality can have a direct impact on how humans think, behave, see others, and feel emotions (Hassabis et al., 2014; Peeters et al., 2006). This makes personality an informative factor to consider when examining differences between individuals and various patterns of behavior and cognition (Connelly et al., 2018). It is no surprise then that this topic has gained some attention in the field of HRI. The various ways that personality has been examined in HRI can be broadly categorized into three perspectives: (1) how humans' personalities impact HRI, (2) how humans' perceptions of robots' personalities impact HRI, and (3) how the degree of similarity or difference between humans' personalities and their perceptions of robots' personalities impacts HRI.

10.4.2 How Does a Human's Personality Impact HRI?

Humans' personalities have been shown to have direct and indirect impacts on a range of outcomes in HRI. In particular, two recent qualitative reviews of the literature on personality identified 20 sets of outcomes across the HRI literature that were examined in reference to personality (Robert, 2018; Robert Jr et al., 2020). These outcomes can be more concisely grouped into seven

overarching categories. These are proximetrics (e.g., distance, touch), attitudes, anthropomorphism, performance, trust, emotional response, and acceptance. Across each of these categories, with the exception of anthropomorphism, results have been somewhat mixed (Robert, 2018; Robert Jr et al., 2020). In particular, studies examining how personality impacts proximetrics, attitudes, trust, emotional response, and acceptance each found significant results at some points and non-significant results at others.

Fortunately, subsequent work in this domain has accounted for one set of these mixed results. In particular, a meta-analysis based on a review of the literature examined the impact of personality on acceptance. This study (Esterwood et al., 2021b) showed that personality appeared to significantly impact acceptance. This impact, however, was not without moderators, and the authors identified other factors as influential. Additional analysis on other outcomes with mixed results, however, has not been conducted given the relatively small number of studies examining these relationships. Future work is therefore needed, but for the time being human personality does appear to directly impact humans' acceptance of robots and anthropomorphism.

10.4.3 How Does a Robot's Personality Impact HRI?

While the impact of humans' personalities in HRI has received the majority of attention in the HRI literature, there is growing research focused on a robot's perceived personality as well. A robot's personality can be defined in a similar manner to that of humans – i.e., a set of distinctive patterns of behavior – however, it is important to note that a robot's personality does not arise from the robot itself. Instead, robot's personalities emerge through the observations, interactions, and expectations of humans of a given robot (Esterwood et al., 2022; Tay et al., 2014). In this way, a robot's personality is predominantly shaped by human's own perceptions which, in turn, are influenced by multitudinous other factors. Although a comprehensive understanding of these factors is currently lacking in the literature, certain common aspects of a robot's design, such as voice cues, gestures, facial expressions, posture, and body movement, appear to play a role (Lee et al., 2006; Mileounis et al., 2015). It should be noted that these factors have not been thoroughly assessed regarding personality traits beyond extroversion. Nonetheless, studies have shown that faster speech, higher pitch, increased volume of speech, and more dynamic and rapid movements are often associated with perceptions of a robot as extroverted, while the opposite characteristics tend to make the robot appear introverted (Lee et al., 2006; Mileounis et al., 2015).

Regardless of these factors, the impact of robot's personalities – however they are manifested – appear significant. Indeed, recent reviews of the literature have uncovered 31 outcomes associated with robot personality in HRI. These can be grouped into three categories: attitudes (i.e., perceptions of robots), acceptance (i.e., intention to use robots), and interaction quality (i.e., enjoyment, fun, perceived control) (Robert, 2018; Robert Jr et al., 2020). Findings from this literature have generally shown mixed results, but most studies indicate that for each of these groups of outcomes robot personality is likely an important factor to consider. This is especially the case for acceptance because this outcome has benefited from meta-analysis. In particular, humans' willingness to accept robots was found to be significantly impacted by the type of personality humans see the robot as possessing (Esterwood et al., 2022). Taken together, this literature highlights that robot personality is an important consideration when designing robots and that this aspect of HRI can influence a range of outcomes.

10.4.4 What Impact Does Similarity Between Humans' and Robots' Personalities Have in HRI?

In the previous two sections, we introduced the effects that both human personality and robot personality have on HRI. Human and robot personalities, however, also interact. Namely, the similarities or differences between human personality and robot personality can have implications for a

range of outcomes. To date, the outcomes examined in the literature have included humans' perceptions of the quality of a robot, their perceptions of the quality of their interactions with the robot, their perceptions of the robot's personality, and their willingness to accept robots (Robert, 2018; Robert Jr et al., 2020).

Generally, each of these outcomes has produced both significant and non-significant results, making firm conclusions difficult. Acceptance, however, has been examined via meta-analysis, which has allowed for closer examination of the impacts of similarity in personality on acceptance. In particular, results indicated that personality similarity between humans and robots appears to have a significant and positive relationship with acceptance (Esterwood et al., 2021a). These findings further highlight how both humans' and robots' personalities are important considerations when designing effective HRIs.

10.5 EXPLANATION IN HRI

The field of explainable artificial intelligence (XAI) has gained renewed interest as researchers seek to improve transparency, interpretability, and understandability of AI systems in response to ethical concerns and a lack of trust (Angwin et al., 2022; Miller, 2019; Stubbs et al., 2007). Research suggests that transparent decision-making processes and algorithms are crucial to building trust in AI systems (Hayes and Shah, 2017; Luo et al., 2022; Zhang et al., 2021).

Explanations, which refer to the reasoning or logic behind actions, can provide vital information that justifies the decisions of intelligence agents, ultimately leading to more trusting and efficient interactions (Hayes and Shah, 2017; Miller, 2019; Zhang et al., 2021). Explanations are necessary for various AI applications, including HRI. Sharing expectations about behavior and intentions between humans and robots is crucial for building trust and understanding the robot's behavior, which can improve communication and collaboration effectiveness (Setchi et al., 2020).

10.5.1 Explainable AI in HRI

Strategies that focus on conveying intention and resolving ambiguity in verbal interactions contribute to the effectiveness of HRI and collaboration. Explainable AI can help individuals without in-depth knowledge of AI understand, predict, and ultimately trust AI systems, as well as identify and address any potential issues or errors in the robot's decision-making processes (Miller, 2019). To achieve explainable AI in HRI, some techniques include natural language generation to explain the robot's actions (Bisk et al., 2016; McDonald, 2010; Tellex et al., 2020), visual explanations (Edmonds et al., 2019; Mishra et al., 2022), and causal reasoning to demonstrate the logic behind the robot's decisions (Alaieri and Vellino, 2016; Erdem et al., 2011; Mota et al., 2021).

Providing explanations about a robotic system's intention, state, capability, and upcoming actions can significantly aid users in developing an accurate mental model of the system (Kulesza et al., 2013). This model helps users continuously understand the explanation, anticipate future actions, and take necessary precautions when unforeseen circumstances arise (Naujoks et al., 2017a). Using a theory of mind to guide robot behaviors and information sharing, as demonstrated by Devin and Alami (2016), led to increased collaboration efficiency when knowledge was communicated appropriately. Additionally, robots capable of identifying and communicating relevant details about their behavior and reasoning make better teammates than those lacking these capabilities because users can then better understand the system and the reasoning behind its actions, as noted by Körber et al. (2018).

Explanations can help clarify the responsibilities of users and robot systems, particularly as robots become more autonomous and individuals tend to attribute blame to the robot instead of themselves or their coworkers (Kim and Hinds, 2006). Through cooperative perception, explanations can demonstrate that both parties are partners by explaining how the system operates and clarifying what users are expected to do. An understanding of whether it is the users or the AI system that determines the system's behavior enables more effective interaction with the system

(Naujoks et al., 2017a; Stanton and Young, 1998). This understanding is particularly important when the robot teammate performs an unexpected action because an explanation can help to increase the perceived responsibility of both parties.

Explanations play a significant role in shaping human attitudes toward and acceptance of robots. For instance, Ambsdorf et al. (2022) designed a HRI scenario to investigate the impact of robots using XAI to explain their actions and found that robots providing reasoning about their actions were perceived as more human-like and lively than those simply announcing their actions. Rational explanations can also improve trust and performance among users who are unfamiliar with a task (Schaffer et al., 2019). Example-based explanations have been shown to provide a better understanding of the system and have a positive impact on trust (Cai et al., 2019). This assertion was supported by Chiou et al. (2022), who demonstrated that increasing awareness of the purpose, process, and performance of robot teammates can help humans retain situational awareness.

10.5.2 Explanation Timing, Content, and Modality and HRI

The role of explanations in HRIs is pivotal, contributing significantly to the acceptance of robots. Robots are often engineered to mimic human intelligence and physicality. Their capacity to provide explanations assists humans in developing precise mental models. These explanations supply crucial information justifying the robot's behaviors, thereby enabling humans to comprehend and anticipate the robot's actions. An increasing number of studies are investigating the influence of robot explanations on behavioral and attitudinal outcomes (Du et al., 2019; Forster et al., 2017; Han et al., 2021; Lettl and Schulte, 2013; Lyons et al., 2023; Koo et al., 2016; Körber et al., 2018; Naujoks et al., 2017b; Ruijten et al., 2018; Shen et al., 2020, Stange and Kopp, 2021; Zhu and Williams, 2020). Notably, three substantial areas of robot explanations – timing, content, and modality – have been recognized and examined in the research field.

10.5.2.1 Explanation Timing

The timing of explanations, namely the point at which a robot offers explanation, plays a vital role in enhancing the efficacy of such clarifications. One can categorize explanation timing into three groups: pre-action (explanation offered before action), in situ (explanation concurrently with action), and post-action (explanation after action).

Various studies have delved into the impact of pre-action robot explanations. For example, Stange and Kopp (2021) scrutinized the repercussions of a social robot, Pepper, offering proactive self-explanations before versus after engaging in an undesirable behavior. Results suggested that while participants experienced less uncertainty regarding future events, they also felt less in control, and displayed diminished trust and lower intentions of interacting with a robot that proactively explained its actions. Similarly, Zhu and Williams (2020) found no compelling evidence linking proactive explanations to positive outcomes in human–robot team tasks. In fact, they noted potential drawbacks to such explanations, as participants perceived them as verbose and unnecessary. Contrarily, research in the realm of automated vehicles (AV), a subset of mobile robots, indicated that pre-action explanations tend to yield positive outcomes. Koo et al. (2016) discovered that explanations issued 1 second before an AV's action alleviated driver anxiety and heightened their sense of control, preference, and alertness. Consistent with this, Du et al. (2019) observed that explanations given 7 seconds before an AV's action garnered more trust and preference than those offered post-action or not at all, reducing anxiety and workload. Moreover, verbal messages describing an AV's impending action, relayed 7 seconds prior, generated higher levels of trust, anthropomorphism, and usability (Forster et al., 2017). Ruijten et al. (2018) found that an AV supplying pre-action explanations appeared more trustworthy, intelligent, human-like, and likable than those devoid of such clarifications.

The impact of in situ explanations during tasks in human–robot collaborations has also been explored. Han et al. (2021) employed a Rethink Robotics Baxter humanoid robot for a handover task

in HRI, assessing participants' perceptions of the necessity and timing of robot explanations. Their findings demonstrated that participants universally acknowledged the need for robots to provide explanations and emphasized the importance of in situ timing.

Post-action explanations have likewise been subjected to investigation. Du et al. (2019) noted that explanations furnished 1 second after the AV's action resulted in the lowest levels of trust and preference, compared to pre-action explanations or none at all. Körber et al. (2018) conducted a mixed-design study which offered explanations 14 seconds post-action, revealing that while drivers believed they understood the system and the rationale for the AV's action, their trust in and acceptance of AVs did not significantly improve compared to when no explanation was provided. Shen et al. (2020), using AV driving videos, examined which driving scenarios necessitated explanations, and how the requirement for explanation differed according to the situation and driver type. The research identified a correlation between the need for an explanation, the driver type, and the driving scenario, with more aggressive drivers requiring fewer explanations. However, near-crash situations unequivocally demanded clear explanations (Shen et al., 2020).

10.5.2.2 Explanation Content

Explanation content pertains to the information about robotic actions provided to humans, and its influence on human reactions has been a key subject of past research. The content of explanations has been classified into three types: (1) 'what' – the actions executed by the robot (Stange and Kopp, 2021; Miller, 2019; Koo et al., 2015; Wiegand et al., 2019), (2) 'why' - the rationale behind the actions (Lyons et al., 2023; Han et al., 2021; Koo et al., 2016), and (3) 'what+why' – encompassing both actions and the reasons behind them (Koo et al., 2015; Du et al., 2019). Different types of explanation content have demonstrated varied effects on human attitudes and behavior.

'What-only' explanations convey solely the actions taken by the robot (i.e., what it will do or did do). Revealing a robot's intended actions can increase perceived understandability if the robot's activities appear ambiguous to users (Stange and Kopp, 2021). It might also positively influence user perceptions by suggesting that the robot is cognizant of potential misunderstandings and is proactively addressing them (Miller, 2019). However, this type of explanation also has its limitations. A study by Koo et al. (2015), using a fixed-base driving simulator with a realistic AV model, found that 'what-only' explanations led to lower acceptance and poorer driving performance compared to other explanation types ('what+why', 'why-only', and no explanation). 'What-only' explanations were deemed the least acceptable and most hazardous in terms of driving performance.

'Why-only' explanations deliver the logic behind robotic actions. Research by Han et al. (2021) discovered the significance of 'why' explanations (e.g., reasoning behind certain behaviors, failures, disobedience, etc.) in HRIs. Without an explanation for unclear behavior or task incompletion, participants inferred that there were issues with the robot needing resolution. Similarly, Lyons et al. (2023) used an autonomous search robot in an Urban Search and Rescue (USAR) scenario to investigate the role of explanations when the robot deviated from expected behavior. Explanations that emphasized the robot's awareness of the environment and why certain events occurred were found to effectively mitigate decreases in trust and trustworthiness. Moreover, 'why-only' explanations resulted in reduced anxiety and improved trust, preference, and driving performance, compared to other types of explanations in AV domain (Koo et al., 2015; Koo et al., 2016). These explanations help drivers anticipate events, maintain control, and enhance situational awareness. Wiegand et al. (2019) found that presenting 'why-only' explanations, including details about detected object movements and contextual information, significantly improved people's understanding of the situation and their situational awareness.

The 'what+why' explanation encompasses both the action performed by the robot and the logic behind it. Support for the efficacy of 'what+why' explanations has been found in various studies. For instance, Forster et al. (2017) discovered that this type of explanation bolsters trust, anthropomorphism, and usability in mobile robots (i.e., AVs). Additionally, Naujoks et al. (2017a) found that 'what+why' explanations help reduce visual workload, thereby making the automation more

user-friendly as drivers do not need to constantly monitor the system's interface to decipher its intentions and actions. However, it is crucial to note that the influence of 'what + why' explanations on trust in AVs can vary based on factors like the driving event, vehicle action, driving environment, and the perspective of the explanation. In an experiment conducted by Ha et al. (2020), a driving simulator equipped with virtual reality technology was used to evaluate the effect of perceived risk and explanations on trust in AVs. The experiment featured four automated driving environments with varying weather conditions (clear day, snowy night) and speeds (fast, slow), and three explanation conditions: no explanation, 'what + why' explanation without a subject, and 'what + why' explanation from a third-person perspective. The findings demonstrated that both the perceived risk of the driving environment and the type of explanation played a vital role in influencing the impact of 'what + why' explanations on trust in AVs. Interestingly, in low-risk perceptions, third-person explanations were most effective in building trust. However, as the perceived risk escalated, the efficacy of third-person explanations diminished, and providing no explanation proved to be the most effective approach.

10.5.2.3 Explanation Modality

The communication approach adopted by robots to relay information to passengers and operators is referred to as their 'modality', which significantly affects both user experience and human perceptions. A modality can be understood as an independent channel through which automation and humans can exchange sensory data (Karray et al., 2008). Research on robot explanations has primarily employed two modalities: auditory and visual. Auditory explanations are typically delivered through a robot or robotic platform using a neutral tone (Lettl and Schulte, 2013; Du et al., 2019; Körber et al., 2018; Koo et al., 2015, 2016; Ruijten et al., 2018; Naujoks et al., 2017b). On the other hand, visual explanations often take the form of text-based natural language processing and annotations (Harbers et al., 2011; Wang et al., 2016), motion or light cues (Baraka et al., 2016; Anjomshoae et al., 2019), or graphical representations and images (Chen et al., 2018; Lim and Dey, 2011) integrated into the user interface to provide explanatory information. In the realm of HRI, expressive motions and lights have been identified as the most effective means of communicating the robot's internal state (Baraka et al., 2016; Anjomshoae et al., 2019).In the automated vehicle domain, while the influence of modality on explanations provided by level 4 and higher AVs is still under investigation, prior research has examined the effectiveness of alert modality in levels 1–3 driving automation, with a particular focus on vehicle display design. Studies generally indicate that auditory modality is a superior choice to visual modality, due to its less distracting nature and its enhanced ability to direct attention compared to visual warnings (Bernsen and Dybkjær, 2001; Cao et al., 2010; Wickens, 2008). This makes it a preferable choice for issuing warnings and promptly conveying potential danger levels (Wheatley and Hurwitz, 2001; Bernsen and Dybkjær, 2001; Cao et al., 2010; Wickens, 2008). However, auditory information has been found to trigger higher levels of perceived annoyance and surprise in drivers, leading to increased stress, delayed reactions, and incorrect responses (Nees et al., 2016; Dingus et al., 1997). In contrast, visual modalities such as icons displayed on a head-up display, boost perceptions of ease-of-use, transparency, and satisfaction (Du et al., 2021; Avetisyan et al., 2022). Visual warnings like texts and icons also support continuous awareness of the surrounding environment and require shorter recognition times for urgency compared to auditory warnings (Politis et al., 2015).

10.6 EVALUATION METRICS IN HRI

Research in the field of HRI has examined a wide breadth of outcomes. As a result, discussing all possible outcomes and the measures for each is outside of the scope of this section. Some outcomes in the HRI literature, however, are especially common. In particular, a recent examination of metrics and methods used in HRI (Zimmerman et al., 2022) as well as a range of systematic reviews and meta-analyses (Esterwood et al., 2021b; Hancock et al., 2011; Naneva et al., 2020;

Roesler et al., 2021) have uncovered a handful of common metrics and measures in HRI. These outcomes can be categorized into two different groups, namely, human-directed measurements and robot-directed measurements. For human-directed measurements, the most popular outcomes across this literature include trust (Hancock et al., 2011), acceptance (Naneva et al., 2020), workload (Prewett et al., 2009), human personality (Esterwood et al., 2022; Robert, 2018; Robert Jr et al., 2020), and attitude (Naneva et al., 2020). For robot-based measurements, the most common measures include anthropomorphism (Roesler et al., 2021), and robot personality (Esterwood et al., 2022; Robert, 2018; Robert Jr et al., 2020). While a complete account of each of the measures used for each of these outcomes is largely absent, some of these outcomes have received special attention. This allows us to examine their associated measures at a high level.

10.6.1 Common Human-Related Measurements

10.6.1.1 Trust

Of the outcomes common to HRI, trust measurements have perhaps received the most attention, with lively debate over the most suitable scales to use, when, and why being common throughout the literature (Chita-Tegmark et al., 2021; Kessler et al., 2017). Exacerbating this debate and the challenges with measuring trust in HRI in general is the frequency of custom instruments (Zimmerman et al., 2022). Although such scales are not inherently invalid and offer many benefits in the form of flexibility, they do limit the reproducibility of results because many of these measures are not fully reported (Zimmerman et al., 2022). Independent of these custom instruments, however, several common measures—referred to as "named surveys"—do exist (Zimmerman et al., 2022).

These named surveys include the Human–Robot Trust questionnaire (Schaefer, 2013), the Madsen and Gregor Measure of Human-Computer Trust (Madsen and Gregor, 2000), the Multi-Dimensional Measure of Trust (MDMT) (Ullman and Malle, 2018), and the Trust in Automation Scale (Jian et al., 2000). In addition, trust has been measured via the Muir Trust Scale (Muir and Moray, 1996), the Mayer Trust Scale (Mayer et al., 1995), and the Lee and See Trust Scale (Lee and See, 2004). Each of these scales purports to measure trust, but each varies in its respective approaches, with some emphasizing different aspects of trust than others. This divergence in measures likely stems from general disagreement in definitions of trust and the degree to which a robot can be considered like a human or instead akin to automation.

Recently, there has been increasing research attention on developing computational trust models, with several notable works (Xu and Dudek, 2015, Guo and Yang, 2021, Soh et al., 2020). Please refer to Section 10.2.3 for details.

10.6.1.2 Attitude

In addition to trust, a common outcome in the HRI literature is that of attitude. Generally, attitude can be considered as an overall construct or divided into various sub-components (Breckler, 1984; Naneva et al., 2020). Measures of attitude as an overall construct have largely relied on the Negative Attitudes toward Robots Scale (NARS) (Nomura et al., 2004), with additional measures such as the Robot Anxiety Scale (Nomura et al., 2008), Attitude toward Technology Scale (Chuttur, 2009), the Ezer Analogies Measure (Ezer, 2008), and the Attitudes toward Working with Robots scale (AWRO) (Robert, 2021) emerging as popular alternatives (Zimmerman et al., 2022).

When considering attitude at a sub-component level, the majority of studies in HRI focus on the affective sub-component of attitudes (Naneva et al., 2020). This has been measured most frequently via the two sub-scales (NARS-S1 and NARS-S2) of NARS (Naneva et al., 2020; Nomura et al., 2004; Zimmerman et al., 2022). Less common alternatives, however, include sub-scales from other measures such as the likeability component of the Godspeed Questionnaire Series (Bartneck et al., 2009) and – as with trust – myriad custom measures (Naneva et al., 2020).

Finally, the cognitive sub-dimension of attitude has largely been measured via one specific sub-scale of NARS (NARS-S2) (Naneva et al., 2020). Various other sub-scales included in measures

of acceptance, however, have also been used (Naneva et al., 2020). Specifically, sub-components of the Almere model of robot acceptance (Heerink et al., 2010) and the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al., 2003) have each been deployed to measure cognitive attitudes in an HRI context (Conti et al., 2017; Shin and Choo, 2011; Tay et al., 2014).

10.6.1.3 Human Personality

Personality has been increasingly examined in HRI (Robert, 2018; Robert Jr et al., 2020). As mentioned in Section 10.3, the most common measure of personality has historically been the Big Five index of personality traits. Unsurprisingly, this index has also been widely used in the context of HRI (Esterwood et al., 2021b; Robert, 2018; Robert Jr et al., 2020; Santamaria and Nathan-Roberts, 2017; Zimmerman et al., 2022). The Big Five index has been commonly used to measure extraversion, neuroticism, openness, agreeableness, and conscientiousness, but most commonly studies have used the index to measure extraversion exclusively (Esterwood and Robert, 2020; Santamaria and Nathan-Roberts, 2017).

While the Big Five personality index remains dominant, it is far from the only approach taken to measure personality. Indeed, measures such as the Myers-Briggs test, the NEO Personality Inventory (Costa and McCrae, 1992), and the Eysenck Personality Inventory (EPI) (Eysenck and Eysenck, 1975) have received attention (Esterwood and Robert, 2020; Santamaria and Nathan-Roberts, 2017). Each of these tests, however, relies on certain assumptions about what a personality is and how it can be used to predict behavior. In particular, these measures of personality exclusively take the trait-based approach to personality. This is but one of many approaches, and criticisms and critiques of the exclusive use of this approach have been presented across the personality psychology literature (McMartin, 2016). Such alternatives, however, have not fully manifested in the HRI domain, and personality remains measured mostly through the Big Five index and, by extension, via trait-based approaches to personality.

10.6.1.4 Workload

Finally, workload is another common outcome measured in the HRI literature (Prewett et al., 2010; Zimmerman et al., 2022). This outcome differs from the others because it is measured in a more standardized fashion. Specifically, workload has consistently been measured across studies with the NASA TLX questionnaire (Hart and Staveland, 1988). In addition, this measure has remained fairly unmodified across studies (Zimmerman et al., 2022). This is unusual because other measures have almost exclusively received some form of modification (Zimmerman et al., 2022). These modifications are mostly minimal alterations to wording to suit the study's design and subjects. As such, the NASA TLX is as close to a standard measure of an outcome as the HRI literature has seen. That is not to say, however, that the NASA TLX is the only measure of workload available to HRI researchers. Indeed, prior work has sought to create HRI-specific measures of workload stemming from the NASA TLX (Yagoda, 2010). This work, however, positions its measure as an accompaniment to the NASA TLX rather than a replacement.

10.6.2 Common Robot-Related Measures

10.6.2.1 Anthropomorphism

Another common outcome in the HRI literature is that of anthropomorphism. This outcome has been most commonly assessed through a measure named the Godspeed (Bartneck et al., 2009; Mara et al., 2022; Roesler et al., 2021; Zimmerman et al., 2022). Godspeed has long been the standard not only for measuring anthropomorphism itself but also, to a lesser extent, for assessing humans' attitudes toward robots (Roesler et al., 2021). The Godspeed, however, is not without limitations, and recent critiques have highlighted several major shortcomings of this measure (Carpinella et al., 2017; Ho and MacDorman, 2010; Kühne and Peter, 2022; Roesler et al., 2021). In particular, the Godspeed appears to suffer from confounded effects, poor factor loading, high

correlation between dimensions, and issues with semantic differentiation in response formats (Ho and MacDorman, 2010).

In response to this criticism, recent work has re-conceptualized anthropomorphism in HRI to more clearly divide this concept from animacy and social presence (Kühne and Peter, 2022). Specifically, such re-conceptualizations have adopted a more multidimensional approach based on the theory of mind. In particular, new dimensions separate from those in the Godspeed have been proposed (Kühne and Peter, 2022). Such approaches, however, are in their infancy. As a result, validation and the establishment of formalized scales—much less their widespread adoption—have not fully emerged.

10.6.3 Robot Personality

The measurement of robot personality parallels that of human personality as measures of robot personality often rely on the trait-based approach to personality as exemplified via the Big Five personality index. Also similar to human personality measures, however, this research has also predominantly concentrated on extroversion, neglecting the exploration of other traits within the Big Five (Esterwood et al., 2022; Santamaria and Nathan-Roberts, 2017). Consequently, a comprehensive understanding of robot personality as a multifaceted construct remains limited, creating an exciting prospect for further investigation and advancement in this field. Exacerbating this issue is the lack of uniformity –and in some cases direct measurements—of robot personalities. Specifically, the HRI literature on this topic to date has diverged greatly in how they opt to measure robot personality (Esterwood et al., 2022). For example, of the seven studies examining extroversion identified by Esterwood et al., (2022) none used the same measurement instruments weakening the capability of future studies to build upon their results. Ultimately, more standardization in measures is clearly needed in regard to robot personality and future work should endeavor not only to use more consistent measures but also to report said measures in full.

10.7 CONCLUSION

In this chapter, we reflected on the HRI literature by examining trust, personality, explanation, and evaluation metrics. Each of these areas represents some of the most influential areas in the study of HRI. Yet, much work remains to deepen our understanding of HRI, with significant implications for scholarship and practice. As robots continue to advance, their use is expected to spread across various spheres of human life. Accordingly, the study of HRI will become increasingly important for our society. In closing, we hope this chapter propels research toward the next step in advancing our understanding of this area.

REFERENCES

Akash, K., Polson, K., Reid, T., and Jain, N. (2019). Improving human-machine collaboration through transparency-based feedback - part I: Human trust and workload model. *IFAC-PapersOnLine*, 51(34):315–321.

Alaieri, F., & Vellino, A. (2016). Ethical Decision Making in Robots: Autonomy, Trust and Responsibility: Autonomy Trust and Responsibility. In Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, November 1–3, 2016 Proceedings 8 (pp. 159–168). Springer International Publishing.

Alarcon, G. M., Capiola, A., and Pfahler, M. D. (2021). The role of human personality on trust in human-robot interaction. In *Trust in Human-Robot Interaction* (pp. 159–178). Elsevier, Cambridge, MA.

Allport, G. W. (1937). *Personality: A Psychological Interpretation*. Henry Holt and Company, New York.

Ambsdorf, J., Munir, A., Wei, Y., Degkwitz, K., Harms, H. M., Stannek, S., Ahrens, K., Becker, D., Strahl, E., and Weber, T. (2022). Explain yourself! Effects of explanations in human-robot interaction. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 393–400). IEEE.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In Kirsten Martin (ed.) *Ethics of Data and Analytics* (pp. 254–264). Auerbach Publications, Boca Raton, FL.

Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13-17, 2019* (pp. 1078–1088). International Foundation for Autonomous Agents and Multiagent Systems.

Avetisyan, L., Ayoub, J., and Zhou, F. (2022). Investigating explanations in conditional and highly automated driving: The effects of situation awareness and modality. *Transportation Research Part F: Traffic Psychology and Behaviour*, 89:456–466.

Azevedo-Sa, H., Jayaraman, S. K., Esterwood, C. T., Yang, X. J., Robert Jr, L. P., and Tilbury, D. M. (2021a). Real-time estimation of drivers' trust in automated driving systems. *International Journal of Social Robotics*, 13(8):1911–1927.

Azevedo-Sa, H., Yang, X. J., Robert, L. P., and Tilbury, D. M. (2021b). A unified bidirectional model for natural and artificial trust in human-robot collaboration. *IEEE Robotics and Automation Letters*, 6(3):5913–5920.

Baker, A. L., Phillips, E. K., Ullman, D., and Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(4):1–30.

Baraka, K., Paiva, A., and Veloso, M. (2016). Expressive lights for revealing mobile service robot state. In *Robot 2015: Second Iberian Robotics Conference: Advances in Robotics, Volume 1* (pp. 107–119). Springer International Publishing.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1:71–81.

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, J. J., Fleeson, W., Grafton, Ben Jayawickereme, E., Kurzius, E., MacLeod, C., Miller, L., Read, S. J., Roberts, B., Robinson, M. D., Wood, D., Wrzus, C., and Mottus, R. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality*, 31(5):503–528.

Bernsen, N. O. and Dybkjær, L. (2001). Exploring natural interaction in the car. In *CLASS Workshop on Natural Interactivity and Intelligent Interactive Information Representation* (Vol. 2, No. 1).

Bhat, S., Lyons, J. B., Shi, C., and Yang, X. J. (2022). Clustering trust dynamics in a human-robot sequential decision-making task. *IEEE Robotics and Automation Letters*, 7(4):8815–8822.

Bisk, Y., Yuret, D., and Marcu, D. (2016). Natural language communication with robots. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational linguistics: Human Language Technologies* (pp. 751–761).

Bostelman, R., Hong, T., and Marvel, J. (2016). Survey of research for performance measurement of mobile manipulators. *Journal of Research of the National Institute of Standards and Technology*, 121:342.

Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology*, 47(6):1191.

Cai, C. J., Jongejan, J., and Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 258–262).

Cao, Y., Mahr, A., Castronovo, S., Theune, M., Stahl, C., and Müller, C. A. (2010). Local danger warnings for drivers: The effect of modality and level of assistance on driver reaction. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (pp. 239–248).

Carpinella, C. M., Wyman, A. B., Perez, M. A., and Stroessner, S. J. (2017). The robotic social attributes scale (rosas) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction* (pp. 254–262).

Cha, E., Kim, Y., Fong, T., and Mataric, M. J. (2018). A survey of nonverbal signaling methods for non-humanoid robots. *Foundations and Trends® in Robotics*, 6(4):211–323.

Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., and Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3), 259–282.

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., and Srinivasa, S. (2020). Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(2):1–23.

Chiou, E. K., Demir, M., Buchanan, V., Corral, C. C., Endsley, M. R., Lematta, G. J., Cooke, N. J., and McNeese, N. J. (2022). Towards human-robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task. *International Journal of Social Robotics*, 14:1117–1136.

Chita-Tegmark, M., Law, T., Rabb, N., and Scheutz, M. (2021). Can you trust your trust measure? In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 92–100).

Chuttur, M. (2009). Overview of the technology acceptance model: origins, developments and future directions. *Sprouts: Working Papers on Information Systems.* 37(9):290.

Coeckelbergh, M. (2011). Humans, animals, and robots: A phenomenological approach to human-robot relations. *International Journal of Social Robotics*, 3:197–204.

Colquitt, J. A. and Salam, S. C. (2012). Foster trust through ability, benevolence, and integrity. In *Locke, Edwin (ed.) Handbook of Principles of Organizational Behavior: Indispensable Knowledge for Evidence-Based Management* (pp. 389–404). John Wiley & Sons, Inc., New York City.

Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, 92(4):909.

Connelly, B. S., Ones, D. S., and Hülsheger, U. R. (2018). Personality in industrial, work and organizational psychology: Theory, measurement and application. In Ones, D. S., Anderson, N., Viswesvaran, C., and Sinangil, H. K. (Eds.) *The SAGE Handbook of Industrial, Work and Organizational Psychology* (Vol. 1, pp. 320–365). SAGE Inc., Thousand Oaks, CA.

Conti, D., Di Nuovo, S., Buono, S., and Di Nuovo, A. (2017). Robots in education and care of children with developmental disabilities: A study on acceptance by experienced and future professionals. *International Journal of Social Robotics*, 9:51–62.

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., and Paiva, A. (2018). Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (pp. 507–513).

Costa, A., Ferrin, D., and Fulmer, C. (2018). Trust at work. In Ones, D. S., Anderson, N., Viswesvaran, C., and Sinangil, H. K. (Eds.) *The Sage Handbook of Industrial, Work & Organizational Psychology* (pp. 435–467). SAGE Inc., Thousand Oaks, CA.

Costa, P. T. and McCrae, R. R. (1992). Normal personality assessment in clinical practice: The neo personality inventory. *Psychological Assessment*, 4(1):5.

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., and Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, 12(2):459–478.

Devin, S. and Alami, R. (2016). An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 319–326). IEEE.

Dingus, T. A., McGehee, D. V., Manakkal, N., Jahns, S. K., Carney, C., and Hankey, J. M. (1997). Human factors field evaluation of automotive headway maintenance/collision warning devices. *Human Factors*, 39(2):216–229.

Do Quang, H., Manh, T. N., Manh, C. N., Tien, D. P., Van, M. T., Tien, K. N., and Duc, D. N. (2020). An approach to design navigation system for omnidirectional mobile robot based on ROS. *International Journal of Mechanical Engineering and Robotics Research*, 9(11):1502–1508.

Driskell, J. E., Goodwin, G. F., Salas, E., and O'Shea, P. G. (2006). What makes a good team player? Personality and team effectiveness. *Group Dynamics: Theory, Research, and Practice*, 10(4):249.

Du, N., Zhou, F., Tilbury, D., Robert, L. P., and Yang, X. J. (2021). Designing alert systems in takeover transitions: The effects of display information and modality. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 173–180), Leeds, UK: ACM.

Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y. N., Lu, H., and Zhu, S.-C. (2019). A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, 4(37):eaay4663.

Erdem, E., Haspalamutgil, K., Palaz, C., Patoglu, V., and Uras, T. (2011). Combining high-level causal reasoning with low-level geometric reasoning and motion planning for robotic manipulation. In *2011 IEEE International Conference on Robotics and Automation* (pp. 4575–4581). IEEE.

Esterwood, C. and Robert, L. P. (2020). Personality in healthcare human robot interaction (h-hri) a literature review and brief critique. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (pp. 87–95).

Esterwood, C. and Robert, L. P. (2021). Do you still trust me? Human-robot trust repair strategies. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (pp. 183–188). IEEE.

Esterwood, C. and Robert, L. P. (2022a). Having the right attitude: How attitude impacts trust repair in human-robot interaction. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 332–341). IEEE.

Esterwood, C. and Robert, L. P. (2022b). A literature review of trust repair in hri. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1641–1646). IEEE.

Esterwood, C. and Robert, L. P. (2023a). Three strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior*, 142:107658.

Esterwood, C., Essenmacher, K., Yang, H., Zeng, F., and Robert, L. P. (2021a). Birds of a feather flock together: But do humans and robots? A meta-analysis of human and robot personality matching. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (pp. 343–348). IEEE.

Esterwood, C., Essenmacher, K., Yang, H., Zeng, F., and Robert, L. P. (2021b). A meta-analysis of human personality and robot acceptance in human-robot interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–18).

Esterwood, C., Essenmacher, K., Yang, H., Zeng, F., and Robert, L. P. (2022). A personable robot: Meta-analysis of robot personality and human acceptance. *IEEE Robotics and Automation Letters*, 7(3):6918–6925.

Esterwood, C. and Robert, L. (2023b). The warehouse robot interaction sim: An open-source HRI research platform. In *2023 ACM/IEEE International Conference on Human-Robot Interaction*.

Esterwood, C., and Robert, L. P. (2023c). The theory of mind and human–robot trust repair. *Scientific Reports*, 13(1), 9877.

Eysenck, H. J. and Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire (Junior & Adult)*. Hodder and Stoughton Educational, London.

Ezenyilimba, A., Wong, M., Hehr, A., Demir, M., Wolff, A., Chiou, E., and Cooke, N. (2022). Impact of transparency and explanations on trust and situation awareness in human-robot teams. *Journal of Cognitive Engineering and Decision Making*, 17(1):75–93.

Ezer, N. (2008). *Is a Robot an Appliance, Teammate, or Friend? Age-Related Differences in Expectations of and Attitudes towards Personal Home-Based Robots*. PhD thesis, Georgia Institute of Technology.

Forster, Y., Naujoks, F., and Neukum, A. (2017). Increasing anthropomorphism and trust in automated driving functions by adding speech output. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (pp. 365–372).

Georgiades, C., German, A., Hogue, A., Liu, H., Prahacs, C., Ripsman, A., Sim, R., Torres, L.-A., Zhang, P., Buehler, M., et al. (2004). Aqua: An aquatic walking robot. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)* (Vol. 4, pp. 3525–3531). IEEE.

Geravand, M., Werner, C., Hauer, K., and Peer, A. (2016). An integrated decision-making approach for adaptive shared control of mobility assistance robots. *International Journal of Social Robotics*, 8:631–648.

Gillespie, N., Lockey, S., Hornsey, M., and Okimoto, T. (2021). Trust repair: A multilevel framework. In Gillespie, N., Fulmer. C., Lewicki, R., (Eds.) *Understanding Trust in Organizations* (pp. 143–176). Routledge, Oxfordshire.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, 4(1):26.

Grover, S. L., Hasel, M. C., Manville, C., and Serrano-Archimi, C. (2014). Follower reactions to leader trust violations: A grounded theory of violation types, likelihood of recovery, and recovery process. *European Management Journal*, 32(5):689–702.

Guinness, D., Muehlbradt, A., Szafir, D., and Kane, S. K. (2019). Robographics: Dynamic tactile graphics powered by mobile robots. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 318–328).

Guo, Y. and Yang, X. J. (2021). Modeling and predicting trust dynamics in human-robot teaming: A Bayesian inference approach. *International Journal of Social Robotics*, 13:1899–1909.

Guo, Y., Shi, C., and Yang, X. J. (2021). Reverse psychology in trust-aware human-robot interaction. *IEEE Robotics and Automation Letters*, 6(3):4851–4858.

Guo, Y., Yang, X. J., and Shi, C. (2023a). TIP: A trust inference and propagation model in multi-human multi-robot teams. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, HRI '23* (pp. 639–643), New York: Association for Computing Machinery.

Guo, Y., Yang, X. J., and Shi, C. (2023b). Enabling team of teams: A trust inference and propagation (TIP) model in multi-human multi-robot teams. In *Proceedings of Robotics: Science and Systems*. Daegu, South Korea.

Ha, T., Kim, S., Seo, D., and Lee, S. (2020). Effects of explanation types and perceived risk on trust in autonomous vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, 73:271–280.

Han, Z., Phillips, E., and Yanco, H. A. (2021). The need for verbal robot explanations and how people would like a robot to explain itself. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(4), 1–42.

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

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., and Szalma, J. L. (2021). Evolving trust in robots: Specification through sequential and comparative meta-analyses. *Human Factors*, 63(7):1196–1229.

Harbers, M., van den Bosch, K., and Meyer, J. J. C. (2011, January). A theoretical framework for explaining agent behavior. In *SIMULTECH* (pp. 228–231).

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In Hancock, P., and Meshkati, N,. (Eds.) *Advances in Psychology* (Vol. 52, pp. 139–183). Elsevier, Amsterdam.

Haslam, N. (2007). Trait psychology. In Haslam, N. (Ed.) *Introduction to Personality and Intelligence* (pp. 17–45). London: Sage Publications Ltd.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., and Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, 24(8):1979–1987.

Hayes, B. and Shah, J. A. (2017). Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17* (pp. 303–312). New York: Association for Computing Machinery.

Heerink, M., Kröse, B., Evers, V., et al. (2010). Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model. *Int J of Soc Robotics*, 2, 361–375.

Hirai, K., Hirose, M., Haikawa, Y., and Takenaka, T. (1998). The development of Honda humanoid robot. In *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No. 98CH36146)* (Vol. 2, pp. 1321–1326). IEEE.

Ho, C.-C. and MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, 26(6):1508–1518.

Hu, W.-L., Akash, K., Jain, N., and Reid, T. (2016). Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine*, 49(32):48–53.

Ishida, T., Kuroki, Y., Yamaguchi, J., Fujita, M., and Doi, T. T. (2001). Motion entertainment by a small humanoid robot based on open-r. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)* (Vol. 2, pp. 1079–1086). IEEE.

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71.

John, O. P., Naumann, L. P., and Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of Personality: Theory and Research* (3rd ed., pp. 114–158). New York: Guilford Press.

Kamide, H., Mae, Y., Takubo, T., Ohara, K., and Arai, T. (2014). Direct comparison of psychological evaluation between virtual and real humanoids: Personal space and subjective impressions. *International Journal of Human-Computer Studies*, 72(5):451–459.

Karray, F., Alemzadeh, M., Abou Saleh, J., and Arab, M. N. (2008). Human-computer interaction: Overview on state of the art. *International Journal on smart Sensing and Intelligent Systems*, 1(1):137–159.

Kessler, T. T., Larios, C., Walker, T., Yerdon, V., and Hancock, P. (2017). A comparison of trust measures in human-robot interaction scenarios. In *Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016, Walt Disney World®, Florida, USA* (pp. 353–364). Springer.

Kim, P., Ferrin, D., Cooper, C., and Dirks, K. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of Applied Psychology*, 89(1):104.

Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication* (pp. 80–85). IEEE.

Kim, W., Kim, N., Lyons, J. B., and Nam, C. S. (2020). Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Applied Ergonomics*, 85:103056.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., and Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 9(4):269–275.

Koo, J., Shin, D., Steinert, M., and Leifer, L. (2016). Understanding driver responses to voice alerts of autonomous car operations. *International Journal of Vehicle Design*, 70(4):377–392.

Körber, M., Prasch, L., and Bengler, K. (2018). Why do i have to drive now? Post hoc explanations of takeover requests. *Human Factors*, 60(3):305–323.

Kox, E. S., Kerstholt, J. H., Hueting, T. F., and De Vries, P. W. (2021). Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, 35(2):30.

Kühne, R. and Peter, J. (2022). Anthropomorphism in human-robot interactions: A multidimensional conceptualization. *Communication Theory*, 33(1), 42–52.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing* (pp. 3–10). IEEE.

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

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

Lee, K. M. and Nass, C. (2003). Designing social presence of social actors in human computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 289–296). ACM.

Lee, K. M., Peng, W., Jin, S. A., & Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. Journal of *Communication*, 56(4):754–772.

Lettl, B. and Schulte, A. (2013). Self-explanation capability for cognitive agents on-board of UCAVs to improve cooperation in a manned-unmanned fighter team. In *AIAA Infotech@ Aerospace (I@ A) Conference* (p. 4898).

Lewicki, R. J. and Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior*, 4:287–313.

Lewicki, R. J., Polin, B., and Lount Jr, R. B. (2016). An exploration of the structure of effective apologies. *Negotiation and Conflict Management Research*, 9(2):177–196.

Li, N., Barrick, M. R., Zimmerman, R. D., and Chiaburu, D. S. (2014). Retaining the productive employee: The role of personality. *Academy of Management Annals*, 8(1):347–395.

Liang, N. and Nejat, G. (2022). A meta-analysis on remote HRI and in-person HRI: What is a socially assistive robot to do? *Sensors*, 22(19):7155.

Lim, B. Y. and Dey, A. K. (2011, August). Design of an intelligible mobile context-aware application. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (pp. 157–166).

Long, J. H., Schumacher, J., Livingston, N., and Kemp, M. (2006). Four flippers or two? Tetrapodal swimming with an aquatic robot. *Bioinspiration & Biomimetics*, 1(1):20.

Lu, S., Zhang, M. Y., Ersal, T., and Yang, X. J. (2019). Workload management in teleoperation of unmanned ground vehicles: Effects of a delay compensation aid on human operators' workload and teleoperation performance. *International Journal of Human-Computer Interaction*, 35(19):1820–1830.

Lu, Y. (2020). *Detecting and Overcoming Trust Miscalibration in Real Time Using an Eye-tracking Based Technique*. PhD thesis, University of Michigan.

Luo, R., Du, N., and Yang, X. J. (2022). Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time. *International Journal of Human-Computer Interaction*, 38(18–20):1962–1971.

Lyons, J. B., aldin Hamdan, I., and Vo, T. Q. (2023). Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior*, 138:107473.

Madsen, M. and Gregor, S. (2000). Measuring human-computer trust. In *11th Australasian Conference on Information Systems* (Vol. 53, pp. 6–8). Citeseer.

Manzey, D., Reichenbach, J., and Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6(1):57–87.

Mara, M., Appel, M., and Gnambs, T. (2022). Human-like robots and the uncanny valley: A meta-analysis of user responses based on the Godspeed scales. *Zeitschrift für Psychologie*, 230(1):33.

Mara, M., Stein, J.-P., Latoschik, M. E., Lugrin, B., Schreiner, C., Hostettler, R., and Appel, M. (2021). User responses to a humanoid robot observed in real life, virtual reality, 3d and 2d. *Frontiers in Psychology*, 12:633178.

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

McCrae, R. R. and Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, 52(5):509.

McCrae, R. R. and Costa Jr, P. T. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, and L. A. Pervin (Eds.), *Handbook of Personality: Theory and Research* (3rd ed., pp. 159–181). The Guilford Press.

McDonald, D. D. (2010). Natural language generation. In Indurkhya, N., and Damerau, F., (Eds) *Handbook of Natural Language Processing* (Vol. 2, pp. 121–144). London.

McMartin, J. (2016). *Personality Psychology: A Student-Centered Approach*. Sage Publications, Thousand Oaks, CA.

Michaud, F., Clavet, A., Lachiver, G., and Lucas, M. (2000). Designing toy robots to help autistic children an open design project for electrical and computer engineering education. In *2000 Annual Conference* (pp. 5–205).

Mileounis, A., Cuijpers, R.H., Barakova, E.I. (2015). Creating Robots with Personality: The Effect of Personality on Social Intelligence. In: *Artificial Computation in Biology and Medicine*. *IWINAC 2015.* Lecture Notes in Computer Science (vol. 9107). Springer, Cham.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Millo, F., Gesualdo, M., Fraboni, F., and Giusino, D. (2021). Human likeness in robots: Differences between industrial and non-industrial robots. In *Proceedings of the 32nd European Conference on Cognitive Ergonomics* (pp. 1–5).

Mishra, A., Soni, U., Huang, J., and Bryan, C. (2022). Why? Why not? When? Visual explanations of agent behaviour in reinforcement learning. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)* (pp. 111–120). IEEE.

Moray, N., Inagaki, T., and Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology Applied*, 6(1):44–58.

Mota, T., Sridharan, M., and Leonardis, A. (2021). Integrated commonsense reasoning and deep learning for transparent decision making in robotics. *SN Computer Science*, 2(4):242.

Moyle, W., Jones, C., Cooke, M., O'Dwyer, S., Sung, B., and Drummond, S. (2013). Social robots helping people with dementia: Assessing efficacy of social robots in the nursing home environment. In *2013 6th International Conference on Human System Interactions (HSI)* (pp. 608–613). IEEE.

Muir, B. M. and Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460.

Naneva, S., Sarda Gou, M., Webb, T. L., and Prescott, T. J. (2020). A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *International Journal of Social Robotics*, 12(6):1179–1201.

Naujoks, F., Forster, Y., Wiedemann, K., and Neukum, A. (2017a). A human-machine interface for cooperative highly automated driving. In N. A. Stanton, S. Landry, G. Di Bucchianico, and A. Vallicelli, (Eds.), *Advances in Human Aspects of Transportation* (pp. 585–595), Cham: Springer International Publishing.

Naujoks, F., Forster, Y., Wiedemann, K., and Neukum, A. (2017b). Improving usefulness of automated driving by lowering primary task interference through HMI design. *Journal of Advanced Transportation*, 2017:6105087.

Nees, M. A., Helbein, B., and Porter, A. (2016). Speech auditory alerts promote memory for alerted events in a video-simulated self-driving car ride. *Human Factors*, 58(3):416–426.

Nomura, T., Kanda, T., Suzuki, T., and Kato, K. (2004). Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)* (pp. 35–40). IEEE.

Nomura, T., Kanda, T., Suzuki, T., and Kato, K. (2008). Prediction of human behavior in human-robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Transactions on Robotics*, 24(2):442–451.

Peeters, M. A., Van Tuijl, H. F., Rutte, C. G., and Reymen, I. M. (2006). Personality and team performance: A meta-analysis. *European Journal of Personality*, 20(5):377–396.

Piazza, C., Grioli, G., Catalano, M., and Bicchi, A. (2019). A century of robotic hands. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:1–32.

Pocius, K. E. (1991). Personality factors in human-computer interaction: A review of the literature. *Computers in Human Behavior*, 7(3):103–135.

Politis, I., Brewster, S., and Pollick, F. (2015). To beep or not to beep? Comparing abstract versus language-based multimodal driver displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15 (pp. 3971–3980), New York: Association for Computing Machinery.

Poon, J. M. (2013). Effects of benevolence, integrity, and ability on trust-in-supervisor. *Employee Relations*, 35(4):396–407.

Prewett, M. S., Johnson, R. C., Saboe, K. N., Elliott, L. R., and Coovert, M. D. (2010). Managing workload in human-robot interaction: A review of empirical studies. *Computers in Human Behavior*, 26(5):840–856.

Prewett, M. S., Saboe, K. N., Johnson, R. C., Coovert, M. D., and Elliott, L. R. (2009). Workload in human-robot interaction: A review of manipulations and outcomes. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 53, pp. 1393–1397). Sage CA: Los Angeles, CA: SAGE Publications.

Raghunath, N., Myers, P., Sanchez, C. A., and Fitter, N. T. (2021). Women are funny: Influence of apparent gender and embodiment in robot comedy. In *Social Robotics: 13th International Conference, ICSR 2021, Singapore, Singapore, November 10-13, 2021*, *Proceedings 13* (pp. 3–13. Springer.

Rawlins, B. (2008). Measuring the relationship between organizational transparency and employee trust. *Public Relations Journal*, 2(2):1–21.

Rhee, J., Parent, D., and Basu, A. (2013). The influence of personality and ability on undergraduate teamwork and team performance. *SpringerPlus*, 2(1):16.

Robert Jr, L. P., Alahmad, R., Esterwood, C., Kim, S., You, S., Zhang, Q., et al. (2020). A review of personality in human-robot interactions. *Foundations and Trends® in Information Systems*, 4(2):107–212.

Robert, L. (2018). Personality in the human robot interaction literature: A review and brief critique. In *Proceedings of the 24th Americas Conference on Information Systems, Aug* (pp. 16–18).

Robert, L. P. (2021). A measurement of attitude toward working with robots (AWRO): A compare and contrast study of AWRO with negative attitude toward robots (NARS). In *Human-Computer Interaction. Interaction Techniques and Novel Applications: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24-29, 2021, Proceedings, Part II 23* (pp. 288–299). Springer.

Robert, L. P., Denis, A. R., and Hung, Y.-T. C. (2009). Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *Journal of Management Information Systems*, 26(2):241–279.

Robinette, P., Howard, A. M., and Wagner, A. R. (2015). Timing is key for robot trust repair. In *International Conference on Social Robotics* (pp. 574–583). Springer.

Roesler, E., Manzey, D., and Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, 6(58):eabj5425.

Roitberg, A., Schneider, D., Djamal, A., Seibold, C., Reiß, S., and Stiefelhagen, R. (2021). Let's play for action: Recognizing activities of daily living by learning from life simulation video games. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 8563–8569). IEEE.

Ruijten, P. A. M., Terken, J. M. B., and Chandramouli, S. N. (2018). Enhancing trust in autonomous vehicles through intelligent user interfaces that mimic human behavior. *Multimodal Technologies and Interaction*, 2(4):62.

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y., and Hancock, P. A. (2011). A model of human-robot trust: Theoretical model development. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 55, pp. 1432–1436). Sage CA: Los Angeles, CA: SAGE Publications.

Santamaria, T. and Nathan-Roberts, D. (2017). Personality measurement and design in human-robot interaction: A systematic and critical review. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, pp. 853–857). Sage CA: Los Angeles, CA: SAGE Publications.

Schaefer, K. (2013). *The Perception and Measurement of Human-Robot Trust*. PhD thesis, University of Central Florida.

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., and H¨ollerer, T. (2019). I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 240–251).

Schneier, M., Schneier, M., and Bostelman, R. (2015). *Literature Review of Mobile Robots for Manufacturing*. US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.

Schweitzer, M. E., Hershey, J. C., and Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and human Decision Processes*, 101(1):1–19.

Sebo, S. S., Krishnamurthi, P., and Scassellati, B. (2019). "I don't believe you": Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 57–65). IEEE.

Setchi, R., Dehkordi, M. B., and Khan, J. S. (2020). Explainable robotics in human-robot interactions. *Procedia Computer Science*, 176:3057–3066.

Shen, Y., Jiang, S., Chen, Y., Yang, E., Jin, X., Fan, Y., and Campbell, K. D. (2020). To explain or not to explain: A study on the necessity of explanations for autonomous vehicles. *arXiv preprint arXiv:2006.11684*.

Shin, D.-H. and Choo, H. (2011). Modeling the acceptance of socially interactive robotics: Social presence in human-robot interaction. *Interaction Studies*, 12(3):430–460.

Soh, H., Xie, Y., Chen, M., and Hsu, D. (2020). Multi-task trust transfer for human-robot interaction. *The International Journal of Robotics Research*, 39(2-3):233–249.

Sørensen, C., De Reuver, M., and Basole, R. C. (2015). Mobile platforms and ecosystems. *Journal of Information Technology, 30:195–197*.

Stange, S. and Kopp, S. (2021, November). Explaining before or after acting? How the timing of self-explanations affects user perception of robot behavior. In *Social Robotics: 13th International Conference, ICSR 2021, Singapore, Singapore, November 10-13, 2021, Proceedings* (pp. 142–153). Cham: Springer International Publishing.

Stanton, N. A. and Young, M. S. (1998). Vehicle automation and driving performance. *Ergonomics*, 41(7):1014–1028.

Stubbs, K., Wettergreen, D., and Hinds, P. (2007). Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems*, 22(2):42–50.

Suzuki, S. (2018). Recent researches on innovative drone technologies in robotics field. *Advanced Robotics*, 32(19):1008–1022.

Tasa, K., Sears, G. J., and Schat, A. C. (2011). Personality and teamwork behavior in context: The cross-level moderating role of collective efficacy. *Journal of Organizational Behavior*, 32(1):65–85.

Tay, B., Jung, Y., and Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior*, 38:75–84.

Tellex, S., Gopalan, N., Kress-Gazit, H., and Matuszek, C. (2020). Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55.

Terada, K., Shamoto, T., Ito, A., and Mei, H. (2007). Reactive movements of non-humanoid robots cause intention attribution in humans. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 3715–3720). IEEE.

Tsui, K. M., Desai, M., Yanco, H. A., and Uhlik, C. (2011). Exploring use cases for telepresence robots. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 11–18.

Ullman, D. and Malle, B. F. (2018). What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 263–264).

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27:425–478.

Völkel, S. T., Schödel, R., Buschek, D., Stachl, C., Au, Q., Bischl, B., Bühner, M., and Hußmann, H. (2019). 2 opportunities and challenges of utilizing personality traits for personalization in HCI. In Augstein, M., Herder, E., and Worndl, W. (Eds.) *Personalized Human-Computer Interaction* (p. 31), De Gruyter, Berlin.

Wang, N., Pynadath, D. V., and Hill, S. G. (2016, May). The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (pp. 997–1005).

Wheatley, D. J. and Hurwitz, H. B. (2001). The use of a multi-modal interface to integrate in-vehicle information presentation. In *Driving Assessment Conference* (Vol. 1, No. 2001). University of Iowa.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, 50(3):449–455.

Wiegand, G., Schmidmaier, M., Weber, T., Liu, Y., and Hussmann, H. (2019). I drive - you trust: Explaining driving behavior of autonomous cars. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19 (pp. 1–6), New York: Association for Computing Machinery.

Xu, A. and Dudek, G. (2015). 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.

Xu, A. and Dudek, G. (2016). Maintaining efficient collaboration with trust-seeking robots. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3312–3319).

Yagoda, R. E. (2010). Development of the human robot interaction workload measurement tool (hri-wm). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54 (pp. 304–308). Sage CA: Los Angeles, CA: Sage Publications.

Yang, X. J., Guo, Y., and Schemanske, C. (2023). From trust to trust dynamics: Combining empirical and computational approaches to model and predict trust dynamics in human-autonomy interaction. In V. G. Duffy, S. J. Landry, J. D. Lee, and N. A. Stanton, (Eds.), *Human-Automation Interaction: Transportation* (pp. 253–265). Springer, Cham, Berlin.

Yang, X. J., Schemanske, C., and Searle, C. (2021). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, 65(5):862–878.

Yang, X. J., Unhelkar, V. V., Li, K., and Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17* (pp. 408–416), ACM, New York.

You, S. and Robert Jr, L. P. (2018). Emotional attachment, performance, and viability in teams collaborating with embodied physical action (EPA) robots. *Journal of the Association for Information Systems*, 19(5):377–407.

You, S., Kim, J.-H., Lee, S., Kamat, V., and Robert Jr, L. P. (2018). Enhancing perceived safety in human-robot collaborative construction using immersive virtual environments. *Automation in Construction*, 96:161–170.

Zanchettin, A. M., Bascetta, L., and Rocco, P. (2013). Achieving humanlike motion: Resolving redundancy for anthropomorphic industrial manipulators. *IEEE Robotics & Automation Magazine*, 20(4):131–138.

Zhang, Q., Yang, X. J., and Robert, L. P. (2021). What and when to explain? A survey of the impact of explanation on attitudes toward adopting automated vehicles. *IEEE Access*, 9:159533–159540.

Zhang, X. (2021). *"Sorry, it was My Fault": Repairing Trust in Human-Robot Interactions*. Thesis, University of Oklahoma.

Zhu, L. and Williams, T. (2020). Effects of proactive explanations by robots on human-robot trust. In *Social Robotics: 12th International Conference, ICSR 2020, Golden, CO, USA, November 14-18, 2020, Proceedings 12* (pp. 85–95). Springer International Publishing.

Zimmerman, M., Bagchi, S., Marvel, J., and Nguyen, V. (2022). An analysis of metrics and methods in research from human-robot interaction conferences, 2015-2021. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 644–648). IEEE.