Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals (in press)

Article *in* The Academy of Management Annals · April 2020

CITATIONS
4

READS
12,786

2 authors:

Ella Glikson
Bar Ilan University
26 PUBLICATIONS **821** CITATIONS
SEE PROFILE

Anita Williams Woolley
Carnegie Mellon University
64 PUBLICATIONS **4,998** CITATIONS
SEE PROFILE

HUMAN TRUST IN ARTIFICIAL INTELLIGENCE:

REVIEW OF EMPIRICAL RESEARCH

Ella Glikson
Bar Ilan University
Ramat Gan, 5290002
Israel
Tel: *[972]54-6466688
Email: ella.glikson@biu.ac.il

Anita Williams Woolley
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA, 15213
Tel: [412] 268-2287
Email: awolley@andrew.cmu.edu

Accepted for publication at the *Academy of Management Annals*, March 2020

HUMAN TRUST IN ARTIFICIAL INTELLIGENCE:

REVIEW OF EMPIRICAL RESEARCH

Abstract

Artificial Intelligence (AI) characterizes a new generation of technologies capable of interacting with the environment and aiming to simulate human intelligence. The success of integrating AI into organizations critically depends on workers' trust in AI technology. This review explains how AI differs from other technologies and presents the existing empirical research on the determinants of human ***trust*** in AI, conducted in multiple disciplines over the last twenty years. Based on the reviewed literature, we identify the form of AI representation (robot, virtual, embedded) and the level of AI's machine intelligence (i.e. its capabilities) as important antecedents to the development of trust and propose a framework that addresses the elements that shape users' cognitive and emotional trust. Our review reveals the important role of AI's tangibility, transparency, reliability and immediacy behaviors in developing cognitive trust, and the role of AI's anthropomorphism specifically for emotional trust. We also note several limitations in the current evidence base, such as diversity of trust measures and over-reliance on short-term, small sample, and experimental studies, where the development of trust is likely to be different than in longer-term, higher-stakes field environments. Based on our review, we suggest the most promising paths for future research.

Artificial Intelligence (AI) represents a highly capable and complex technology that aims to simulate human intelligence. AI sits at the core of what has been termed the "fourth industrial revolution" (Schwab, 2017), distinguished by the shift of agency and control from humans to technology, and thus transforms our previous understanding of human-technology relations. This revolution and its implications highlight new theoretical and empirical questions that need to be addressed by organizational researchers, as AI has the potential to dramatically change the overall workforce structure as well as the way organizations and jobs are designed, decisions are made, and knowledge is managed (Brynjolfsson, Mitchell, & Rock, 2018; Danaher, 2017; Huang & Rust, 2018; Kaplan, 2015; Kellogg, Valentine, & Christin, 2019; Pfeffer, 2018; Wirtz et al., 2018). The exact shape of these changes is still to be determined; this leaves room for an open, multidisciplinary dialogue that should explore human-AI collaboration and further facilitate the way AI is developed. The trust that users develop in AI technology will be central to determining its role in organizations moving forward. Thus, we review the latest empirical research to lay a foundation for understanding the ways humans develop trust in AI.

As development of trust among humans is highly dependent on the physical appearance of the trustee (Cho & Hu, 2009; Duarte, Siegel, & Young, 2012) AI embodiment is likely to be an important consideration in trust development between humans and AI. Researchers have examined AI in a variety of embodiment forms: as a physical robot; a virtual agent or bot; or in forms that are invisible to the user, embedded inside of a computer or other tool. In addition to variance in AI embodiment, researchers examined human trust in AI under different levels of AI machine intelligence, i.e. its capabilities. Higher machine intelligence means more complex technological abilities, which allow AI to produce more autonomous and complex actions (Chen & Barnes, 2014; Hancock et al., 2011). Users are not always aware of the actual technological sophistication of AI;

while in some cases highly intelligent machines are acting in their full capacity, in others, the capability may not be fully manifest in their behavior. For current purposes, we focus on the trust of human users, and thus address the perceived machine intelligence from the users' point of view.

Proceeding from the intersection of research on AI and the extant literature on human trust, we organize our review of the existing literature by considering the physical appearance of AI (i.e., its representation), addressing the level of machine intelligence, and looking at the implications of each for the development of both cognitive and emotional trust (McAllister, 1995). In contrast to the existing reviews and meta-analyses that focused on studies from a specific field, such as robotics (Hancock et al., 2011) or human factors (Hoff & Bashir, 2015; Lee & See, 2004), this review integrates research from different disciplines, providing a comprehensive overview. For each AI representation (robotic, virtual and embedded), we discuss the common dimensions that emerged from our review as relevant for cognitive trust (tangibility, transparency, reliability, task characteristics and immediacy behaviors) and for emotional trust (tangibility, anthropomorphism and immediacy behaviors).

Literature Review Methodology

This review presents the way trust in AI is currently discussed in the literatures of computer science, human-computer interaction, human factors, information systems, robotics, management, marketing, and psychology. Focusing on human trust in AI, we first used the Google Scholar platform, searching for the following key words: artificial intelligence (AI), intelligent agents, agent-human interaction, algorithm aversion, robot-human interaction, intelligent automation, trust in robot, trust in technology. We limited the search to articles published in the last 20 years (between 1999 and 2019) to address the empirical work concomitant with the recent technological

development of AI. We screened articles based on the content, including those relevant to human trust in AI, while excluding descriptions of algorithm/architecture (without reference to trust), or those focusing on trust only among humans. We then followed cross-reference techniques to find more relevant articles. This search revealed approximately 200 peer-reviewed papers and conference proceedings from the fields of Organizational Behavior, Human-Computer Interactions, Robot-Human Interactions, Information Systems, Information Technology, and Engineering. Finally, we used three databases, Business Source Premier, Engineering Village, and PsycINFO, to complete the literature review and, using the same guidelines, added an additional 50 papers based on their content. Out of the mentioned articles, only about 150 have presented empirical research that directly or indirectly addresses human trust in AI. We have also included most recent published review papers that focus on trust in technology or in robotics in more general terms.

In proceeding with our review, we first define AI and then discuss the broader perspective of its integration into organizations and review the concept of trust from a multidisciplinary perspective. We present the empirical research for the three major types of AI representations and consider the intersection with the levels of machine intelligence, first for the development of cognitive and then for emotional trust. Next, we discuss the integration of the research, as well as its implications for organizations, the existing limitations, and directions for future research.

What is Artificial Intelligence?

In management research, Artificial Intelligence (AI) is defined as a new generation of technologies capable of interacting with the environment by (a) gathering information from outside (including from natural language) or from other computer systems; (b) interpreting this

information, recognizing patterns, inducing rules, or predicting events; (c) generating results, answering questions, or giving instructions to other systems; and (d) evaluating the results of their actions and improving their decision systems to achieve specific objectives (Ferràs-Hernández, 2018). The interactional properties of AI make it capable of learning and changing its behavior based on the cues from the environment (Frantz, 2003; Rahwan et al., 2019). As the environment in which AI functions is usually highly complex and partially random, AI's behavior is not deterministic (Danks & London, 2017), and the complex multi-layer process of AI decision-making is generally not transparent. This means that AI's decisions could be difficult to predict, and the logic behind each decision made is often poorly understood.

The futuristic literature assumes AI is a set of algorithms able to perform *all tasks* just as well as, or even better than, humans. However, this type of superintelligence, known as Strong or General AI, does not yet exist, and thus this review is focused on the Weak or Narrow AI that is currently in use (Raj & Seamans, 2019; Russell & Norvig, 1995). Weak AI is based on a variety of technologies that are able to achieve *fragments* of the simulation of human intelligence, such as face recognition. To better understand how AI differs from more traditional technology, it is useful to explain one commonly used component of AI, namely machine learning.

Machine learning is the ability of computers to adjust their behavior based on the data to which they are exposed (Samuel, 1959). This means that having a specific goal, such as minimizing number of misses, and a set of rules that define what is a miss or a hit, will enable computers to adjust their decisions based on their experiences. This learning process requires a large amount of data that can be used for training. When properly trained, AI is able to make accurate decisions with newly presented similar data, and adjust its behavior when necessary (Brynjolfsson &

Mitchell, 2017). However, the training process may introduce unintended bias via the features of the data, the algorithm, or the data-algorithm interaction (Danks & London, 2017).

There are two important assumptions in machine learning: First, while the goal is being established by the programmer (for instance, to minimize misses vs. maximize hits), the specific calculations that lead computers to the decision are based on the data and are mostly unknown. Second, the computers are able to utilize data to a greater extent than humans, and may therefore achieve better results than humans (Brynjolfsson & Mitchell, 2017). For example, consider Arthur Samuel, one of the pioneers of machine learning, who taught a computer program to play checkers (Frantz, 2003). His goal was to teach it to play checkers better than himself, which was not something he could program explicitly. Samuel used a large number of annotated games, with the good moves distinguished from the bad ones, and a Guide to Checkers to adjust the criteria for choosing moves, so that the program would choose those moves thought to be good by checker experts as often as possible. In 1962 Samuel's program beat the checkers champion of the state of Connecticut, who was the fourth in the nation, as it was able to play checkers better than its programmer (McCarthy & Feigenbaum, 1990). Even though the program was applying pre-programmed rules, it learned to play better than its creator, making decisions in a better way than the programmer could, which is a key difference between prior generations of technology (which were limited by the knowledge of the programmer) and AI.

Considering the unique qualities of AI technology, it is important to address the difference between AI and automation, as these terms are often used interchangeably (e.g., Lee & See, 2004). Automation refers to the situation where computers follow pre-programmed rules in order to perform repetitive and monotonic tasks that were previously performed by humans (Parasuraman & Riley, 1997). The behavior produced by traditional automation and its outcomes are pre-

programmed and well understood. Traditional automation is deterministic and does not include any learning processes (e.g., Raj & Seamans, 2019). However, automation can be enabled by AI, which means that machine learning algorithms make the rules that the automated process follows, and they also learn and adjust based on experience and feedback. Consequently, automation plays a role in carrying out the actions determined by an intelligent system, and thus this review includes research that examines trust in intelligent, AI-enabled automation. The specific, technical details of AI are beyond the scope of this review; our focus is on reviewing existing empirical research related to users' perceptions and tendencies to trust or not trust AI technology.

Trust and Integration of AI in Organizations

Early work on the acceptance and use of new technology in organizations tended to focus on user reactions to technological features. For example, the Technology Acceptance Model (TAM; Davis, 1989) emphasizes perceived usefulness (the degree to which a potential user believes that technology will help to perform a task) and ease of use (the perceived utility of the effort to use the new technology) as the main determinants of users' attitudes and behavioral intention to use and accept the system (Davis, 1989). More recent approaches have further added the concept of trust as a predictor of technology acceptance (Ghazizadeh, Lee, & Boyle, 2012; Hoff & Bashir, 2015; Lee & See, 2004; Pavlou, 2003).

A focus on issues of trust allows us to address not only the disuse (a rejection) of technology, but also its misuse (an inappropriate over-reliance on technology), or its abuse (harmful use in order to achieve an individual gain; Parasuraman & Riley, 1997). Trust can predict the level of reliance on technology, while the level of correspondence between user's trust and the technology's capabilities, known as calibration, can influence the actual outcomes of technology use. Low trust in highly capable technology would lead to disuse and high costs in terms of lost

time and work efficiency, as well as possible abuse, whereas high trust in incapable technology would lead to over-trust and misuse, which in turn may cause a breach of safety and other undesirable outcomes (Hoff & Bashir, 2015; Lee & See, 2004).

One of the most cited definitions of trust was suggested by Mayer et al. (1995), who argued that trust is "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, Davis, & Schoorman, 1995). This definition emphasizes a willingness to be vulnerable and the importance of the actions at stake, and does not limit the concept of trust to human-human interaction, allowing us to consider trust with regards to technology, including AI (Wang, Qiu, Kim, & Benbasat, 2016). Though definitions in other literatures incorporate some different assumptions, such as socially embedded properties of human or institutional relations, the conceptualization of trust as a tendency to take a meaningful risk while believing in a high chance of positive outcome is common across different disciplines (Hoff & Bashir, 2015; Rousseau, Sitkin, Burt, & Camerer, 1998).

Trust is particularly relevant to the human-AI relationship because of the perceived risk embedded in human-AI relations, due to the complexity and non-determinism of AI behaviors. AI is also perceived as technology that slowly will take over different types of (currently) human jobs as well as fundamentally transform the structure of organizations (G. F. Davis, 2019). It is still not clear whether low-skilled and low-cost employees, such as frontline service representatives, are at a higher risk of being replaced by AI (Huang & Rust, 2018; Pfeffer, 2018; Wirtz et al., 2018) than knowledge workers and top-level managers that rely on analytical and rational-knowledge processing, and whose high cost makes their replacement financially attractive (Ferràs-Hernández, 2018; Loebbecke & Picot, 2015). In the present, some "human" tasks are already being performed

by AI (Brynjolfsson & Mitchell, 2017). Analyzing tasks across almost 1,000 occupations, Brynjolfsson et al., (2018) found that most occupations in most industries have at least some tasks that could be replaced by AI (suitable for machine learning). However, there is no occupation in which all the current tasks could be replaced (Brynjolfsson et al., 2018). That being said, there is also no disagreement that the labor force will go through a dramatic change, with some jobs disappearing and new jobs being created (Faraj, Pachidi, & Sayegh, 2018).

Trust is a dynamic concept that is prone to changes based on the behavior of the trusted agent (Crisp & Jarvenpaa, 2013; Schoorman, Mayer, & Davis, 2007). Hoff and Bashir (2015) posited that the way trust in technology unfolds differs from the way it develops in humans, due to the common positivity bias toward new technologies (Parasuraman & Manzey, 2010). In contrast to the low trust that exists initially between unfamiliar humans, new technologies may produce unrealistically optimistic beliefs regarding their abilities and functionality (Dzindolet et al., 2003). Thus, while trust in humans generally increases with time through frequent interactions, the trust in technology decreases with time, based on encounters with errors and malfunctions (Madhavan & Wiegmann, 2007). However, the opposite also could be true when it comes to AI. Pointing out the widespread skepticism associated with the immaturity of existing AI (Hengstler, Enkel, & Duelli, 2016), and the difficulties associated with the acceptance of new technologies (Leonardi, 2009), some researchers suggest that an initially low level of trust from an initial encounter may increase following a direct interaction (Ullman & Malle, 2017). In this review we address the dynamic nature of trust by discussing the trajectory over which trust develops for users interacting with different AI representations. We also examine the AI features that facilitate the development of trust, such as tangibility, transparency, reliability, and immediacy behaviors, the

context of the task being performed, and the role of machine intelligence in moderating the impact of experience with AI on human trust.

Much of the extant organizational research has considered trust to be a cognitive construct that involves rational evaluation of the trustee and situational features (Schoorman et al., 2007). However, trust might also be influenced by irrational factors, such as emotions and mood (Komiak & Benbasat, 2006). McAllister (1995) referred to the latter as emotion-driven or affect-based trust, suggesting that in interpersonal relationships, people develop social connections that provide support and comfort—in addition to cognitive trust that is based on perceptions of trustee reliance and competence. The emotional trust between human co-workers differs from cognitive trust not only in its antecedents, but also in its behavioral consequences (Jones & George, 1998; Ng & Chua, 2006). Recognizing the differences between trust in humans and trust in technology, Hoff and Bashir (2015) argued that for understanding the adoption of a complex new technology it is essential to address emotion-driven trust. Use of unknown sophisticated technology, such as AI, suggests a need for a "leap of faith" (Hoff & Bashir, 2015; Lee & See, 2004) and trust in processes that cannot be directly observed or cognitively understood. Therefore, in this review we address empirical research that examined both cognitive and emotional trust in AI.

Building Cognitive Trust in AI

Research on human trust attests to the importance of an object's representation and tangibility for establishing trust, and the extant research on AI also supports this notion (Krämer, Lucas, Schmitt, & Gratch, 2017; Lee, Jung, Kim, & Kim, 2006; Li, 2015). However, research that addresses more than one type of AI representation (such as physical robot vs. virtual bot) is rare, as different disciplines tend to focus on a specific type of representation. At the same time, similar representations are often studied by several disciplines. For instance, robotics and human factors

researchers mostly study trust in robots, while researchers in information systems, marketing and human-computer interactions study trust in recommendation agents. Thus, organizing a review by field of study would not be helpful. Taking a user-centered approach and embracing the physical embodiment notion that addresses the physical representation of AI (Lee et al., 2006), we organize the review based on the way AI is presented to human users, separating between AI-enabled robots, AI-enabled virtual agents, and embedded AI. Reviewing the research across disciplines that are relevant to each representation, we directly address the role of AI representation.

As argued previously, trust could be both cognitive (based on rational thinking) and emotional (based on affect; McAllister, 1995), and as these types of trust might differ in their antecedents, we discuss their development separately.[1] Therefore, we start by discussing the dimensions influencing cognitive trust within each physical representation, such as tangibility, transparency, and reliability, task characteristics and immediacy behaviors with the latter reflecting the special interactive abilities of AI. In each section, we present an overview of the findings, the trajectory of trust and review extant research on each dimension relevant to each type of technology representation (robotic, virtual or embedded).

When researchers examine cognitive trust in AI, they measure it as a function of whether users are willing to take factual information or advice and act on it, as well as whether they see the technology as helpful, competent, or useful. Based on prior research that examined trust in

[1] It is important to note that in many cases researchers did not make a clear distinction regarding the type of trust in AI they study; we inferred this distinction based on the context of the study, trust antecedents and trust measures. When the studied mechanism was based on affect or emotions, we categorized it as emotional trust, and when it was based on cognition or rationality, as cognitive trust.

technology (Hancock et al., 2011; Hoff & Bashir, 2015; Lee & See, 2004), we organize the review of cognitive trust in AI by addressing the dimensions that were found to influence trust. We start each section by discussing the *trajectory of trust* as is evident in the reviewed research and the specific role of AI *tangibility*, i.e. its capability of being perceived or touched, in developing trust. An overview of our conclusions can be found in Table 1 and Figure 1.

------Insert Figure 1 and Table 1 about here-----

Transparency reflects the level to which the underlying operating rules and inner logics of the technology is apparent to the users and is considered to be critical for developing trust in new technology (Hoff & Bashir, 2015). It is more problematic for AI than other technologies, especially when methods such as deep learning are involved. An important aspect of transparency includes different types of explanations regarding how AI works or why a specific decision was made that are understandable to users, even when they have little technical knowledge. The review focuses on the studied implications of these explanations for cognitive trust.

Reliability, or exhibiting the same and expected behavior over time, is also critical to technology trustworthiness (Hoff & Bashir, 2015). In the case of AI, reliability is often difficult to assess, especially in the context of high machine intelligence, as learning from data can lead technology to exhibit different behavior, even if the underlying objective function remains the same. As our review reveals, the relationship between AI reliability and trust are less straightforward in high versus low intelligence technologies and varies across AI representations.

Technologies are believed to be more efficient in some tasks than in others, and therefore *task characteristics* related to the work the technology is performing, such as whether it deals with largely technical versus interpersonal judgments, could be an important antecedent for cognitive

trust in AI (Hancock et al., 2011). High machine intelligence significantly improves the performance of AI in traditional technology-related tasks, but also increases the range of tasks that could be performed by technology. As the range of tasks AI can perform keeps growing, the role of task characteristics in developing cognitive trust becomes more complex and less stable.

High machine intelligence that allows the technology to interact with the environment and be responsive to users has introduced a variety of AI's *immediacy behaviors*. Immediacy has been defined as the degree of perceived physical and/or psychological closeness between people (Mehrabian, 1967). Immediacy behaviors refer to socially-oriented gestures intended to increase interpersonal closeness, such as proactivity, active listening, and responsiveness. These behaviors are perceived as signs of machine intelligence and influence cognitive trust by raising the expectations of high-quality performance and positive experience during mutual work.

Cognitive trust in robotic AI

AI-enabled robots may have a variety of functions and capabilities as well as different mechanical or human-like representations; they could be physically present or distantly located and perform mechanical or socially-oriented tasks. Based on the extant research, this review focuses on interactions with physically-present robots, addressing remote robots only with regards to transparency and task characteristics. The initial trust in robotic AI is relatively low, therefore factors such as reliability, transparency and characteristics of the task could play an important role in developing trust. Nevertheless, it seems that a much more important role is being played by the level of machine intelligence that allows robots to engage in a variety of behaviors that increase immediacy, such as responsiveness to users.

Trust trajectory

Extant research addressing the trajectory of human trust in robotic AI suggests that the initial trust starts at a low level and develops over time, as depicted in Figure 1. This means that trust in robotic AI develops in a manner that is similar to trust in humans and increases following direct interaction (Haring et al., 2016; Ullman & Malle, 2017). For instance, Waytz et al. (2014) found that participants who drove a car that was partially autonomous reported higher trust in its abilities than participants without such experience. Bartneck et al. (2006) noted that a short interaction with a robotic pet significantly improved the overall attitude toward robots. Even for children, a hands-on experience with an automated robot was found to increase trust more than other activities, such as watching a video that explained the robot's abilities (Rossi et al., 2018).

Additional evidence for low initial trust in robotic AI is observed in both lab and field studies. For instance, Ullman and Malle (2017) tested the way people trust a small robotic vehicle capable of generating and adjusting its paths. The researchers compared a condition in which the robot was autonomous and performed required adjustments without any human involvement to a condition in which, to activate the adjustment, participants had to push a button. They found that participants reported higher cognitive trust in the robot they controlled. Furthermore, following this controlled experience of involvement, participants expressed significantly higher trust in potential future robots (Ullman & Malle, 2016, 2017).

In field studies, robots were also treated with great suspicion and low initial trust. Importantly, low initial trust does not necessarily lead to a lack of use but can be translated into a misuse of technology, especially in real-life situations. For example, Andrist et al. (2016) conducted a field study in which they analyzed human-robot interactions in a lobby of an office building. The robot's goal was to provide directions to different building areas, such as elevators,

upon users' requests. Analyzing the videotaped interactions across several days, researchers reported that 81% of the interactions were playful, with no real intent to get directions. In the next stage, the robot was programmed to respond to a playful approach, such as laughter, with a direct question regarding the intent of the user. Researchers found that only 15% of users admitted to misusing the robot, while others insisted that they were truly asking for the robot's assistance (Andrist et al., 2016). Despite users denying being playful, their actual use of a robot has the potential to teach about its abilities and build trust.

As the level of machine intelligence of robotic AI increases, it becomes capable of engaging in greater immediacy behaviors. As a result of the robot's engagement in such human-like behaviors, users' initial trust steeply increases and leads to greater user compliance with the robot's directives. While users may perceive this compliance as reciprocity, the observed tendency to cede control to a robot, even when the latter demonstrates erroneous function, is worrisome. We will discuss the issue of compliance despite erroneous robot's behavior in the section on reliability.

Tangibility

One of the main factors that is known to influence initial trust is the robot's actual tangible physical presence (for a review see Li, 2015). For instance, Bainbridge et al. (2011) compared a physically present robot and an identical 2D image presented on a screen. They found that participants were quicker to respond to the physical robot. Furthermore, testing trust by examining compliance with an unusual request, the researchers found that participants were more compliant with a physical robot than its 2D representation.

Shinozawa et al. (2005) made similar comparisons, using the willingness to accept a robot's advice as a behavioral measure of trust. They also found higher trust in the physically-present robot's advice; however, this was only for tasks presented in a physical space. When the task was presented on a screen, there was no difference between the robot and the on-screen image conditions. These latter findings suggest that the fit between the AI representation and the presentation of the task may influence human trust, decreasing the positive impact of physical presence for tasks that are completed online.

Looije, Neerincx and Cnossen, (2010) also compared a physically-present robot and its 2D virtual representation and found that the physically present robot was significantly more trusted than its virtual representation. However, problems with the smoothness of the robot's movement harmed its ability to react to the participants' voices and drove the participants to perceive the 2D virtual representation as having a more social personality—being more friendly and kind. As responsive behavior is more easily created in an animated virtual agent than in a robot, it is important to further evaluate the relative effect of AI's physical presence in comparison to virtual AI responsiveness on trust. Based on the existing research, it seems that tangibility is more important to trust than responsiveness, but in long-term interactions the prosocial behaviors could play a more significant role.

Additional evidence for the positive effect of physical presence can be found in a study performed by Cormier et al. (2013). Modeled after Milgram's compliance experiment, researchers asked participants to perform an extremely boring document-sorting task in the presence of a robot that was only able to voice such phrases as "please continue doing the task." The robot had significantly less effect on participants' behavior than a human facilitator, yet 48% of participants

followed the robot's request and continued the task, while openly voicing their boredom and dissatisfaction.

Existing research shows that it is not only physical presence that influences cognitive trust, but also a robot's physical appearance. For instance, researchers found that robot appearance is often interpreted by users as signaling the level of robot intelligence, and even influences moral judgments, so that human-like robots are expected to make human-like moral decisions, in contrast to robots with a mechanical appearance (Malle et al., 2016). Interestingly, human-likeness is not always associated with perceptions of higher intelligence. Carlson et al. (2015) examined the impact of a cooperation-oriented team activity on perceptions of a team member robot. They found that the team-building activity improved the perception of the robot's anthropomorphism, but not the perception of the robot's level of intelligence.

Transparency

Despite an extensive focus on the importance of algorithm transparency for cognitive trust, the existing research on the role of transparency for trust in robotic AI is very limited and mostly relates to robots that work in remote locations, focusing on the need for shared awareness (Chen & Barnes, 2014; Stubbs, Wettergreen, & Hinds, 2007). Extant research provides only general support for the positive effect of constant information flow on cognitive trust in robots (Sanders et al., 2014). Future research must address this gap and test the effect of transparency for both remote and collocated robots.

Reliability

An increase in trust following interaction usually suggests highly reliable performance. Interestingly, we could find only a few studies that examined the direct impact of reliability of AI-

enabled robots on trust. For instance, Robinette, Howard, and Wagner (2017) showed in several studies that in high-risk situations participants lost trust in the advice of a robot that made a mistake. However, different factors may significantly moderate the relationship between a robot's failure and subsequent human trust. For instance, Desai et al. (2013) examined the moderating effect of the timing of a robot's failure. They found that early drops in reliability lowered real-time trust more than later drops. Freedy et al. (2007) reported similar results connecting the early failures to the first impression regarding the robot's capabilities. However, comparing three different levels of the robot's reliability across several trials, researchers also found that experience with a low reliability robot increased trust, even though the robot was consistently failing. Working with inconsistent reliability (i.e., a medium level of reliability) was more confusing to the participants and their trust in this condition was lower than in the low reliability condition. Some level of caution should be exercised in generalizing these findings, however, as the study had only twelve participants. It is important for ongoing research to examine whether these results would replicate in a large sample, and if reliability is more important for trust in robotic AI than the quality of its performance.

A high level of machine intelligence allows AI-enabled robots not only to assist humans in a variety of tasks, but also to engage in managerial activities that exert some control over users' behavior, such as task allocation, task instructions, or guidance. Reviewing studies in which robots played a semi-managerial role reveals that a robot's reliability could play a less important role for human trust and compliance than previously assumed. For instance, Salem et al. (2015) found that people followed a human-like robot's instructions even when they witnessed its erroneous behavior. The faulty behavior had a significant impact on participants' perceptions of reliability and trust; however, these perceptions did not translate into reduced compliance. Researchers

reported that 91% of participants followed all robot instructions, even the unreasonable ones, such as pouring juice on a plant and disclosing a personal password. Similarly, Robinette et al. (2015; 2016) found that people tend to follow a robot's lead in an emergency situation, even when its prior behavior indicates a serious malfunction.

Despite the obvious limitations regarding the external validity of these laboratory-based studies, including the short-term interactions and low actual risk involved (even in the emergency scenario it is not clear to what extent the participant actually felt danger), the tendency to follow a highly intelligent robotic AI, even when its' actions are questionable, is worrisome and requires more research. We will further discuss the aspects of robots' erroneous behavior in the section on emotional trust.

Task characteristics

Looking into the effect of task characteristics on trusting behavior, Gaudiello et al. (2016) measured the extent to which participants were willing to change their answers following a robot's advice. Researchers used functional questions, such as an evaluation of objects' weight, color and sound, and questions of a social nature regarding the importance of different objects in social scenery, such as a public pool. All cases presented uncertain situations in which any answer could be correct, while the human-like robot provided advice that was always the opposite of the participant's opinion. Results indicate that people conform with the robot more readily on functional issues than on social issues; however, the effect size was small, and the overall rate of accepting the robot's advice on any issue was low (significantly lower than 50%).

In the area of team interpersonal dynamics, Martelaro et al. (2015) tested the willingness of participants to cooperate with a robot's intervention into team conflict. The robot was designed

to intervene when one of team members (a confederate) became very rude. The task of the robot was to restore appropriate communication and avoid deterioration of the conflict. The researchers found, in contrast to what was expected, that the robot's intervention made the existence of conflict more visible to the team members.

Gombolay et al. (2015) tested the effects of shared decision-making authority in human-robot and human-only teams in a manufacturing setting. They found that although people value human teammates more than robotic teammates, they trusted the robot's ability to schedule tasks and manage the workflow. Thus, for a task that requires complex analysis and optimization to reach an effective flow of actions, participants tend to cede their control and authority to the robot, demonstrating high trust. As robots gain more capabilities related to facilitating or even managing team dynamics, it is important to note that for human users to trust and accept a robot's actions, the task allocated to the robot should be well-matched to its actual abilities.

These studies demonstrate the importance of the task for developing cognitive trust in a robot, emphasizing the benefits of tasks that involve complex calculations and technological abilities versus tasks with social features. These findings are consistent with the MABA-HABA ("Machines Are Better At vs. Humans Are Better At") framework that signifies the actions in which machines have significant advantages over humans, such as objective calculations (Bradshaw, Feltovich, & Johnson, 2011; de Winter & Dodou, 2014; M. K. Lee, 2018).

Nevertheless, as a robot's level of machine intelligence increases, so does the ability to demonstrate responsive, prosocial behaviors, and raises the expectation that robots will be able to fill more social roles, such as companionship. Strohkorb-Sebo et al. (2018) examined the effect of a robot's disclosed vulnerability on team dynamics and collaboration. Playing the role of a team facilitator, a humanoid (i.e., human-like) robot was designed to make comments on the team's

progress during a task performance. The robot's tactical expressions were compared with more emotional, encouraging or disappointed expressions. When the robot's behavior was more emotional, team members were more active in reducing tension when they made mistakes and exhibited more trust-building behaviors. As discussed in the previous section with respect to pro-social but low reliability robots, as robotic AI behavior becomes human-like, it becomes easier for users to trust and follow them regardless of the exact task (and the level of reliability). Future research must address the moderating role of task characteristics on the effect of machine intelligence and social behaviors in building cognitive trust, not only in lab studies, but also in field settings. It is possible that at a workplace, in contrast to a lab, trust in a highly intelligent robot would still depend on the nature of the tasks being performed.

Immediacy behaviors

Incorporating higher levels of machine intelligence has enabled robots to react to human presence and speech, creating "social-robots," interactive assistants that are able to serve in roles such as an interactive museum guide, team member, or social companion (Bickmore, Pfeifer, & Schulman, 2011; Hinds, Roberts, & Jones, 2004; Zhang et al., 2010). Interestingly, while paying less attention to robots' reliability, researchers have explored the effect of different robots' behaviors on human trust. Overall, the findings indicate that as the level of machine intelligence increases, users expect robots to be more proactive and adaptive. Behaviors that enhance users' experienced immediacy, such as social gestures, are generally helpful, and their mere presence seems to affect human behavior and bring about compliance with a robot's requests, even when robots exhibit mistakes in their behavior. Faulty behavior can reduce trust but, at least in the short term, compliance continues.

Baraglia et al. (2016) examined two different autonomous forms of robotic behavior—reactive and proactive. The goal of the robot was to assist participants in the performance of a sequence of tasks. The reactive robot provided help only after the participant failed in timely task performance. The proactive robot detected participants' movements, was able to anticipate possible failures, and initiated help before a task was completed. Comparing these conditions and a condition in which the robot was used as a tool (i.e., operating at a low level of intelligence and activated by a human's request for help), researchers found that the participants not only performed better in the proactive than in the reactive robot condition, but also reported that they preferred the proactive robot over the one acting as a tool. Other studies provided similar results. Hoffman and Breazeal (2007) compared teamwork with reactive and proactive robots, where the proactive robot was programmed to anticipate specific behaviors. They found that people liked the proactive robot more and rated it as a more productive team member.

Directly examining the perceptions of trust, De Visser and Parasuraman (2011) compared stable and adaptive levels of robot autonomy. In the stable condition, the robot always provided assistance, needed or not, while in the adaptive condition the help was provided only for difficult tasks. The level of intervention had no effect on participants' performance; however, participants reported higher trust in the adaptively automated robot. Participants appreciated the ability of the robotic assistant to initiate helping behavior in the appropriate situations and reported higher levels of self-confidence and lower levels of workload when working in the adaptive condition. It seems that timing and fit to the situation may facilitate trust, perhaps due to their connection to perceived higher level of robot machine intelligence.

Oistad et al. (2016) examined the effect of a robot's social-oriented behaviors on users' perceptions and physical proximity to the robot during a box moving task. They found that robot's

user-oriented immediacy gestures, such as approaching the user and nodding toward him/her when in proximity, had a positive impact on users' perceptions of robot's anthropomorphism. Furthermore, social gestures decreased the sense of physical risk and participants kept less distance from the higher immediacy robot than the robot that did not demonstrate these behaviors.

In addition to immediacy behaviors executed by robots, researchers addressed some intervening behaviors that can improve human trust. For instance, Carlson et al. (2015) demonstrated that a teambuilding activity increases trust in a robotic team member. In contrast, You and Robert (2019) found that what increased trust in a robot was not a stronger sense of a team, but an activity in which participants assembled their robot. Researchers suggested that the act of robot assembly increased trust and made the participants identify more with their robotic team member. However, the differences in these experiments could also be driven by the assumed machine intelligence of the robots. In the Carlson et al. (2015) experiment, the robot was engaged in a complex search activity and thus had assumed a high machine intelligence; it was in reality operated remotely in a "Wizard of Oz" methodology (i.e., a human was actually controlling the report, unbeknownst to participants), yet participants perceived it as a highly intelligent, autonomously functioning robot. In contrast, in You and Robert's (2019) experiment, the robot had the role of a water carrier, and thus its assumed level of machine intelligence was low. Therefore, it seems that the perceived level of machine intelligence moderates not only the steepness of the trust trajectory, but also the activities and psychological perceptions that lead to cognitive trust.

Cognitive trust in virtual AI

An AI-enabled virtual agent is a representation in which AI has no physical presence, but a distinguished identity, such as a chatbot or an avatar (Ben Mimoun, Poncin, & Garnier, 2012).

Virtual representation may exist on any electronic device, and may possess features such as a face, body, voice, or the ability to text. Despite being already in commercial use, much of the existing empirical research focuses on the aspects of the interface design, paying less attention to such factors as level of machine intelligence.

Trust trajectory

The trust trajectory of virtual AI suggests that high initial trust *decreases* following an interaction (Hoff & Bashir, 2015). This trajectory differs from trust development in robotic AI (see Figure 1) and is evident in both lab and field studies. For instance, De Visser et al. (2017) found that in an initial stage of the experiment, participants trusted the advice of virtual AI more than human advice, yet with time (and following decreasing reliability) this trust decreased much more than the trust in a human adviser. Examining the field evidence of the effect of virtual agents, Ben Mimoun et al. (2012) analyzed their use in commercial websites and found that despite the initial interest in their use, over the years their actual use had significantly decreased. Based on interviews and the fact that this problem was specific to virtual agents, researchers suggested that the lack of calibration between an agent's representation and its actual level of machine intelligence led to customers' frustration and abandonment. Human-like representation of AI may lead to users' expectations of a high-level machine intelligence, which in many cases does not fit the technological reality (Ben Mimoun et al., 2012).

Interestingly, this trajectory can be reversed, with some evidence for low initial trust in virtual AI. Similar to observations made with robotic AI in a field study (Andrist et al., 2016), research found that in some cases, the initial trust in field settings could be low and lead to an agent's misuse and negative behavior. For example, Kopp et al.(2005) analyzed the interaction of museum visitors with a virtual guide. Examining more than 200 conversations, researchers noted

that although most of the visitors tended to greet the virtual agent and 20% tested the system by asking direct and indirect questions regarding its abilities, 11% of the interactions were negative, including insults and abusive and negative language.

Additionally, there is some evidence of relatively low initial trust that increases following an interaction. Wang et al., (2016), who examined different types of recommendation agents, found that the first-hand experience with a reliable recommendation agent increased participants' trust in comparison to third-hand experience. This suggests that when virtual AI has high machine intelligence and is highly functional, similar to the case of robotic AI, direct interaction can increase the initial trust.

To explain the differences of trust trajectories in virtual AI, researchers suggest addressing the calibration between users' expectations and virtual AI performance. Features of virtual AI, such as visualization, and especially anthropomorphism, may significantly increase users' expectations, while the actual level of AI machine intelligence moderates the direction of the trust trajectory. When agents with low machine intelligence are paired with human-like representations, the users are more likely to start with high expectations and experience a trust decrease. By contrast, virtual agents with high machine intelligence can engage in higher immediacy behaviors, which facilitates a positive trust trajectory.

Tangibility

The existing research supports the notion that the visual presence of human-like or animal-like virtual AI agents increases initial cognitive trust in comparison to a lack of visualization. This means that tangibility has a positive effect on cognitive trust in AI, similar to its effect on robotic AI. Examining the effect of a visually present agent, Chattaraman et al. (2014) found that adding

an avatar's picture on a shopping website increased participants' trust and intention to visit the website again. Similarly, Mumm and Mutlu (2011) found that when feedback for a categorization task was produced by an agent (in the form of a robotic picture), participants reported higher intrinsic motivation in contrast to the condition without explicit visualization. However, Wang et al. (2016) argued that visualization mostly influences the emotional trust and not cognitive trust, as it has less impact on the perception of usefulness. In their study, Wang et al. (2016) manipulated the presence of a visual agent and the presence of a detailed explanation of the agent's recommendation and found that the agent's transparency had a greater effect on cognitive trust than adding a visual representation.

Transparency

One way to moderate unrealistically high expectations from users is to provide an explanation regarding virtual AI functionality. Exploring the role of transparency in facilitating trust in AI, Pieters (2011) suggested a distinction between explanation-for-trust and explanation-for-confidence. Pieters argued that confidence could be seen as a reliance on technology without considering alternatives, whereas trust requires comparison between different options. Explanation-for-trust addresses the way the system works, the "how" question, by revealing details of its internal operations. By contrast, explanation-for-confidence makes the user feel comfortable in using the system by providing information about its external communications, explaining "why" an algorithm should be used.

In building on this distinction, Wang and Benbasat (2007) looked at the recommendations of virtual agents and manipulated the transparency of the algorithm by providing explanations about why and how the agent made its decision, and what the alternatives were. Consistent with Pieters (2011), they found that the explanation for how a decision was made influenced consumers'

beliefs in the competence and benevolence of the virtual agent. The transparency regarding the choice (i.e., why something was chosen) influenced only the perceptions of agent benevolence. This is consistent with the extant literature on the effect of explanations on trust for virtual AI (see Xiao and Benbasat, 2007 for review).

An additional way in which transparency could be helpful for establishing trust is when the reliability of the virtual agent is transparent. For instance, Fan et al., (2008) demonstrated that informing participants regarding the actual reliability of a decision-making agent increased participants' trust and improved performance. When the transparent reliability was low, participants better adjusted their decisions, taking into consideration the agent's advice only when appropriate. The ability to know when to use the virtual agent increased the overall trustworthiness of the agent.

Reliability

Reliability plays an important role in users' trust and trusting behaviors in virtual AI, which differs from robotic AI. Moran et al. (2013) examined compliance with voiced agent instructions in a street team-based game and found that compliance was highly dependent on trust in the agent. When agent reliability was compromised, such as when the instructions led to no revelation of new cues, trust decreased, which also decreased the compliance. It seems that experiencing (without advance knowledge) a virtual AI agent's low reliability significantly differs in its effect on trust than simply being aware of the possibility of low reliability. The actual experience decreases the trust, while the transparency regarding AI's possible errors may increase trust.

Looking beyond AI reliability, researchers have suggested the importance of focusing on the consistency/inconsistency between users' expectations and the actual AI performance.

Factoring in the levels of initial trust allows researchers to better predict the effects of virtual AI behavior. This mechanism was found to explain users' trust across different studies (Xiao & Benbasat, 2007). Glass et al. (2008) interviewed users of an office assistant agent and found that correct expectations regarding the agent's performance and capabilities facilitated trust in the agent.

Task characteristics

Virtual AI is perceived as having benefits with respect to technical issues, such as analyzing data, which is similar to robotic AI. Ramchurn et al. (2016) compared human compliance with agent versus human instructions in the context of a response to a disaster, and found that under virtual agent instructions, the performance was better due to the agent's greater capabilities of gathering information and a clearer method of wording instructions.

Testing compliance with a virtual agent, Jiang et al. (2014) found that a virtual agent playing the role of game instructor was highly trusted, as its orders were usually followed (91% compliance). However, this compliance depended on whether the instructions were aligned with team dynamics. When the virtual AI agent's orders required a dramatic change of team dynamic or interfered with accomplishing a different task, the compliance decreased to less than 40%. It is possible that when instructions were interfering with the way the game was played, the players perceived the AI as less intelligent and therefore less trustworthy.

High machine intelligence allows virtual AI to be used to influence interaction between humans. For instance, de Melo, Marsella, and Gratch (2017) found that humans represented by virtual agents led people to act more fairly toward other humans than humans without such representation. Isbister et al. (2000) found that the ability of an AI-driven agent to match safe vs.

unsafe topics for a discussion influenced the cultural perceptions of American and Japanese participants regarding each other and their actual behavior. Safe topics included movies, music and sports, and AI-driven agents who were "safe" would ask questions on the safe topic at any time they would assess a long pause in a conversation. An "unsafe agent" asked questions on issues like politics, religion and money. In a "safe agent" condition, Americans felt more trust in Japanese partners and had a more positive perception about Japanese people in general. Japanese students in the "safe agent" condition found Americans to be more similar to them. Krämer et al. (2017) found that communication with an interacting agent decreases participants' need to engage in social activities, as they sought less human interaction after using AI.

Introducing AI as a team member may also influence the interaction between people in the team, and even alter team cognition and team communication patterns (Demir, McNeese, & Cooke, 2017; Demir et al., 2015). Specifically, Demir et al. (2017) found that human members of teams with an AI "peer" (referred to as a "synthetic member") made significantly fewer information exchanges than teams with only human members.

Immediacy behaviors

High levels of machine intelligence allow virtual AI to enact more immediacy behaviors that increase trust, such as social responsiveness and personalization of the virtual AI agent's reactions. Pro-social virtual AI's behaviors can be translated to perceptions of the agent's personality. Andrews (2012) demonstrated that an agent's pro-social behaviors led participants to perceive a high level of agent agreeableness, which was associated with higher trust in the agent. Komiak and Benbasat (2006) manipulated the level of personalization provided by different recommendation agents, using either personal or general questions asked by the agent. They found that personalization had a significant positive impact on users' cognitive trust.

A virtual agent's persuasion tactics can also be important. Fenster, Zuckerman, and Kraus (2012), found that an agent that provided examples was more influential than an agent that provided justifications and more persuasive than an agent that presented the subject with both examples and justifications. What is particularly intriguing about this finding is that it suggests that the effectiveness of virtual AI persuasion tactics could differ from tactics that typically work well for humans, where using both examples and justifications was found to be more effective.

Cognitive trust in embedded AI

Completely embedded AI is "invisible" to users, which means that it does not have a visual representation or a distinguished identity. It could be embedded in different types of applications, such as a search engine or a GPS map, and users might be not aware of its existence. Assuming users are aware of embedded AI, there is still an important question of what features engender cognitive trust. Similar to virtual AI, cognitive trust in embedded AI differs from robotic AI in that it is more driven by its reliability and transparency. Similar to robotic AI, the perceived level of expertise or machine intelligence also plays an important role in cognitive trust, as well as the type of task involved, as people believe that algorithms are better at calculation tasks than at social tasks. As the level of machine intelligence increases, the contextual and user-centered factors become more important for cognitive trust because it becomes more difficult to assess AI reliability.

Trust trajectory

Research assessing the trajectory of cognitive trust in embedded AI has mostly focused on the way trust in AI changes based on the feedback regarding its accuracy. Many lab-based studies have demonstrated that people tend to exhibit high initial trust in embedded AI as an algorithmic decision-providing software (de Visser et al., 2017; Dietvorst, Simmons, & Massey, 2015;

Manzey, Reichenbach, & Onnasch, 2012). High initial trust tends to decrease as a result of erroneous AI function and the process of trust restoration requires a lot of time.

The few field studies that exist also demonstrate high initial trust in embedded AI. For instance, researchers examining the effect of wearable algorithmic sensors on users' emotions found that users demonstrated a high level of trust in the sensors and that their emotions were significantly influenced by the feedback they received (Hollis et al., 2018). Researchers who studied the cases of Uber and Lyft drivers also reported high initial trust. For instance, Min K. Lee et al. (2015) tested drivers' experience with an AI-enabled management system and found that they perceived the passenger-driver rating system as efficient in establishing basic trust and service attitudes in the ridesharing systems. However, they also found that low levels of transparency lead drivers to social forums, where they could not only make sense of the system, but also gain knowledge about ways of resisting/or abusing it. Möhlmann and Zalmanson (2017) also found that while they kept using the system, Uber drivers did not trust its managerial decisions and engaged in a variety of actions to resist its management, including gaming the system.

There is also evidence of low initial trust in embedded AI, especially in field studies, where the mistrust could be so high that users may refuse using the embedded AI in the first place. However, field studies that assess trust in embedded AI in organizational settings are scarce. Health-care researchers examining how algorithmic decision aids are being used (or not used) by physicians report significant difficulties in acceptance of the technology in medical settings (Linkov et al., 2017; Panella, Marchisio, & Di Stanislao, 2003). In a field experiment on energy use, Alan et al. (2014) found that participants avoided using an algorithm that was designed to help them save on their electricity bills. The refusal to use the technology further prevents the hands-on experience that would increase trust. This could explain why commercial companies are evasive

about their use of embedded AI (e.g. Eslami et al., 2015). Future research must consider the role of trust in AI within organizations to better understand the specific difficulties that need to be tackled in order to facilitate its use.

Tangibility

The embedded nature of AI representation suggests that people may not be aware that they are using an algorithm-enabled application. Currently, research on the impact of embedded AI users' awareness on trust is very limited. In one study, Eslami et al. (2015) surveyed Facebook users and found that more than half of them (62%) were not aware that an algorithm was managing the information that was displayed to them, making decisions on what should be hidden. Learning about the algorithm's way of working changed users' attributions, perceptions, and behaviors, and overall increased their sense of control. Revealing (or hiding) the use of an algorithm may not only raise important ethical questions, but also have a significant impact on users' long-term trust. However, Eslami et al. (2015) found that, despite being unpleasantly surprised or even angry for not being informed about the use of an algorithm, after learning how it worked, participants kept using the platform. Future research needs to explore the limitations for users' trust recovery in such situations, and the true cost and benefit of users' awareness.

Transparency

Looking for ways to overcome the aversion driven by technology error, researchers have examined the role of transparency on cognitive trust. For instance, Mohlemann and Zalmanson (2017) focused on Uber drivers and noted that the lack of algorithm transparency leads drivers to constantly guess and game the system. Lee et al. (2015) reached a similar conclusion.

Dzindolet et al. (2003) used explanations of the rationale behind possible mistakes made by the machine and demonstrated that such explanations had a significant positive effect on trust. Supporting the notion of the positive effect of transparency in developing trust in AI, Chao et al. (2016) found in a survey of over 700 participants that understanding the technological capabilities of AI embedded in a search engine positively correlated with reliance on the technology and belief in the usability and ease of use of the technology. Although this study did not directly assess trust in AI, its results regarding the perceptions of low risk and high reliance suggest a positive relationship between acknowledgement of AI capabilities and trust.

However, not all provided information has a similar effect. Helldin et al. (2013) showed that when drivers of a simulated autonomous vehicle were warned about the situational uncertainty that would lead an algorithm to err, they reported lower trust and were quicker to retake manual control over the car than participants who did not get the warning. Kizilcec (2016) investigated trust in an algorithmic peer-reviewing system and found that when participants' expectations of their outcome were violated, the explanation regarding how the algorithm worked facilitated trust. However, when the explanation included the raw scores in addition to the algorithmic action description, the levels of trust went down. The author suggested that the introduction of additional data was confusing, which undermined the positive effect of the algorithm transparency.

Following Pieter's framework (2011) that differentiates explanations of how algorithms work into "why" versus "how," Cotter, Cho, and Rader, (2017) examined the way Facebook provided explanations about its News Feed algorithm. They found that most of the information targeted the question of "why" this algorithm should be used, rather than "how" it worked, and suggested that such explanation would improve users' confidence in the system, rather than their trust in the system. However, the researchers focused on the company's behavior rather than users'

perceptions and thus the impact of the explanation on users' trust is not certain. The importance of understanding "how" the algorithm works is also evident in studies that allowed users to slightly modify the algorithm (Dietvorst et al., 2016).

The embeddedness of AI can also lead to questions regarding who it is intended to benefit, and thereby undermine trust. Alan et al. (2014) demonstrated that transparency about the actual beneficiary of the decision aid is another important consideration. In a field experiment, participants were asked to test an application that aimed to reduce their electricity expenses. However, participants questioned the true recipients of the benefits of AI: was it them or the electric company (Alan et al., 2014)? These issues are less likely to surface in laboratory studies, as commercial interests are less likely to be involved.

The complexity of AI algorithms rarely allows for full transparency about the basis of its decisions and the tradeoffs it makes (Ananny & Crawford, 2018). However, communication regarding embedded-AI rationale and its actual abilities may significantly improve the calibration of users' expectations regarding AI performance. Better calibration might lower the initial, unrealistically high trust that was observed in lab studies, while improving the recovery of trust in the case of an erroneous outcome, allowing users to build more effective long-term collaboration with the technology (Hoff & Bashir, 2015).

Reliability

Research that tested the levels of trust driven by AI accuracy and failure have revealed a stable pattern, indicating that errors of an embedded-AI are detrimental to cognitive trust. For instance, Dzindolet et al. (2003) tested an automated decision aid and found that errors significantly decreased trust and reliance on the aid. Separating between visibility of an error and

performance feedback across many trials, researchers have demonstrated that the visibility of an error effects trust in a way that is difficult to repair. Similarly, Manzey et al. (2012) found that an erroneous function had a stronger effect on trust than a correct function, and that the trust recovery process was very long. Consequently, researchers have concluded that positive and negative feedback loops are not symmetrical. Dietvorst et al. (2015) demonstrated a similar effect of an erroneous function, referring to it as *algorithm aversion*. Across five studies, researchers found that participants refused to rely on a forecasting model after seeing it err. Participants preferred to rely on a human forecast and not on an algorithm, even when human errors were more severe than algorithm errors.

Task characteristics

Although AI is assumed to perform better on tasks that involve mathematical skills, such as work scheduling, this advantage is not always evident in the empirical studies. For instance, Lee (2018) did not find a significant difference in initial trust between AI and human decision making for analytical tasks. However, for tasks that involve human skills, such as work evaluation, participants demonstrated higher trust in human decisions than in algorithmic decisions.

The subjective value of self-confidence also plays an important role in trust, as people who perceive themselves as more capable than a machine are less trustful and tend to rely less on the technology (Lewandowsky, Mundy, & Tan, 2000). Logg et al. (2018) found that experts used AI-provided advice less than lay participants did, even when ignoring it decreased experts' performance. Researchers explained this finding by referring to evidence of experts being less appreciative of others' advice than non-experts, suggesting that experts tend to rely more on their own opinion.

Immediacy behaviors

Most of the studies considering immediacy behaviors exhibited by robotic and virtual AI that were reviewed indicated a positive impact on trust; however, it seems that such behaviors can highlight the ability of embedded AI to constantly monitor workers and lead to a decrease of trust. In a study of Uber drivers, Mohlemann and Zalmanson (2017) suggested that constant individual performance evaluation and feedback, only possible through constant tracking, violates drivers' sense of autonomy and decreases their trust. Such constant surveillance is perceived as a way of micro-management and conveys a lack of trust from those deploying the AI, which leads to low trust among the drivers.

Lee et al. (2015) suggested an additional explanation of the drivers' decrease of trust. Following a set of interviews, they concluded that a lack of personalization was a key factor that decreased trust. The system lacked consideration of many specific circumstances, such as female drivers rejecting male passengers late at night for safety reasons. Highly intelligent systems should be able to engage in more immediacy behaviors, such as personalization, which can improve the sense of fairness and trust.

Additional support for the effect of personalization could be found in three field studies conducted by Matz et al. (2017). Researchers tested the effects of psychological persuasion on 3.5 million individuals using psychologically-tailored advertising and found that matching the content of persuasive appeals to individuals' psychological characteristics significantly altered participants' behavior as measured by clicks and purchases. Specifically, they found that persuasive appeals that were matched to people's extraversion or openness-to-experience level resulted in up to 40% more clicks and up to 50% more purchases than their mismatching or non-personalized counterparts.

Embedded AI can also produce immediacy behaviors through an activation of nudges or boosts. A nudge (e.g., a default option) is a change in the choice architecture that shifts human behavior by taking advantage of basic cognitive processes and biases (e.g., inertia, procrastination, and loss aversion). In contrast, a boost (e.g., better information) is a change in the choice architecture that shifts behavior by clarifying the direction an individual should move to achieve personal objectives, which is often accomplished by enhancing an individual's decision-making competencies (Camilleri, Cam, & Hoffmann, 2007). Analyzing the possible effects of nudges and boosts generated by intelligent systems, Burr, Cristianini, and Ladyman (2018) suggested that despite the overall agreement that nudges should be used for benefiting users, they might function as coercive and deceptive tools that could redirect user behaviors toward undesirable outcomes. The deceptive nature of nudges requires further research that reflects not only the effect of nudges on trust, but also the ethical aspects of the use of nudges by AI.

Building Emotional Trust in AI

Emotional trust is not commonly addressed in human relations with technology; however, emotions are known to significantly affect human trusting behaviors (Hoff & Bashir, 2015). Furthermore, AI developers specifically target human emotions by manipulating features of AI representations and behaviors. Making a robot or a bot to look or act like a human or a living thing is known to affect users' emotional reactions toward the technology. However, the effect is not always positive, and may also result in negative emotions, such as a sense of eeriness and fear. Even more surprisingly, some researchers have found that people experienced more positive emotions toward an erroneous than a correctly functioning robot (e.g. Mirnig et al., 2017). Therefore, in addition to understanding users' cognitive trust, there is a growing need to understand what and how such features of the technology affect human emotions and emotional trust.

To organize the review of the empirical research on emotional trust in AI, we address the dimensions which were studied the most in this regard, specifically the role of tangibility, anthropomorphism and immediacy behaviors (see Table 2 for an overview). While tangibility and immediacy behaviors were previously discussed with regard to their impact on cognitive trust, anthropomorphism was rarely mentioned. Anthropomorphism, i.e., human-likeness, refers to the perception of technology or an object as having human qualities, such as feelings. These perceptions could be driven by interface features, such as the human-like form of the robot; by behavioral features, such as gaze, node; and by intentional framing, such as giving a robot or bot a human name.

----- Insert Table 2 about here -----

Emotional trust in robotic AI

Tangibility

Robots are known for evoking a variety of emotional reactions in human users, including excitement, but also fear and a sense of eeriness. In contrast to the overall positive effect of tangibility on cognitive trust, its effect on emotional trust is mixed, and might depend on the attitudes of the user. While some people tend to enjoy the physical presence of a robot, others find its tangibility threatening. Focusing on the psychological mechanism that explains humans' predisposition to trust robots, Nomura et al. (2006), studying Japanese students, developed a Negative Attitude toward Robots Scale (NARS) that has been used in many HRI studies. For instance, in a scenario-based study Tussyadiah et al. (2019) found a strong negative correlation between NARS and trusting beliefs regarding functionality, helpfulness and reliability of serving

robots. Bartneck et al., 2006) found that cultural background plays an important role in forming attitudes toward robots, with US participants having the most positive perceptions.

The tangibility of a robot can be influenced by its physical posture. Following Nomura et al.'s (2006) suggestion regarding negative predispositions toward autonomous (i.e., AI-enabled) robots, Obaid, Sandoval and colleagues (2016) examined the physical distance between a human and human-like robot in a task that required physical proximity. The researchers found that people were more willing to approach a sitting robot than a standing one. Users interpreted the posture as a signal of possible physical risk that reduced trusting behavior. The emotional sub-scale of the NARS (negative attitude toward robot scale) was significantly correlated with the kept distance, while users with prior experience were more willing to approach the robot (Obaid et al., 2016b).

At the same time, there is also evidence of the positive effect of robotic tangibility. For instance, Shim and Arkin (2016) showed that elderly participants reported that feedback provided by a robot was more enjoyable, motivating, and trustworthy than one delivered by a computer screen. It seems that when the initial predispositions toward robots are not negative, such as in the case of US users (Bartneck et al., 2006), tangibility may increase emotional trust. Thus, the pre-existing attitudes of users could be an important moderator of the effect of AI agent tangibility on the development of emotional trust.

Anthropomorphism

Anthropomorphism, i.e., human-likeness, is generally thought to have a positive effect on human perceptions and emotions. However, there is also evidence for its negative effect. While examining anthropomorphic robots, some researchers have built on the Uncanny Valley theory (Mori, 1970) which argues that an encounter with an artificial agent that possesses human-like

features leads to an experience of eeriness or a sense of unpleasantness that brings to mind thoughts of mortality (Ho & MacDorman, 2010; Złotowski, Yogeeswaran, & Bartneck, 2017). Research based on these theories examines the negative effects of human-likeness on users' perceptions and emotional trust. Zlotowski and colleagues (2016) examined human interactions with machine-like and human-like robots and found that the machine-like robot was perceived as more empathetic and more trustworthy than the human-like robot, regardless of its positive or negative attitude (Złotowski et al., 2016). Appel et al., (2016) compared participants' perceptions of a robot based on detailed descriptions. They found that when a robot was described as more human-like and having greater agency, it was perceived as uncannier than a less intelligent robot. In the same vein, Zlotowski et al. (2015, 2016) found that a human-like robot induced higher levels of participant anxiety than a machine-like robot.

Interestingly, the studies that reported a negative effect of anthropomorphism on human emotions and emotional trust examined mostly the initial trust, based on a short interaction or description. It seems that gaining an interaction experience with a tangible, human-like robot may decrease a sense of unpleasantness (Złotowski et al., 2015). It is also possible that the mismatch between a robot's appearance and its machine intelligence is a significant source for negative impressions. While it is assumed that human-level machine intelligence (General AI) may evoke emotional discomfort and fear, this type of machine intelligence currently does not exist. In the current state, low emotional trust could be evoked by an anthropomorphic robot that lacks any intelligence, as in the case of Zlotowski et al.'s (2015) study of a human-like robot.

Despite some of the research pointing to the negative effects associated with human-likeness, most of the empirical research focuses on the positive emotions, such as excitement, curiosity and liking, which are the results of interactions with robots. High levels of interest and

acceptance of robots are evident across different populations, including children and the elderly (Jacq et al., 2016; Strohkorb et al., 2016; Zhang et al., 2010). For instance, Zhang et al. (2010) tested different features of a service robot with elderly participants and found that more human-like features of the robot were associated with more emotional trust and the pleasantness of the interaction experience. Furthermore, anthropomorphism had a positive impact on users' physiological parameters, such as heart rate. The positive power of anthropomorphism was also demonstrated by Waytz et al. (2014), who anthropomorphized an autonomous vehicle by giving it a name and a voice. The results indicated that an anthropomorphized car was more trusted and less blamed for errors than one that was simply mechanical.

Immediacy behaviors

In contrast to the human-like appearance, human-like behaviors consistently induce high emotional trust and liking in robotic AI. Bickmore et al. (2013) tested the effectiveness of a robot museum guide and found that its responsiveness had a significant effect on visitors' engagement, learning, enjoyment, and trust. Birnbaum et al. (2016) found that robot responsiveness increased nonverbal approach behaviors such as leaning toward the robot, eye contact and participants' smiling, as well as their willingness to be accompanied by the robot during stressful events. Jung et al. (2013) focused on back-channeling (i.e., interactional cues of active listening, which are mostly non-verbal, such as nodding or moving toward) as an engagement strategy of a robot. They found that this type of behavior displayed by a robot lowered participants' stress and cognitive load.

An intriguing study examined how participants react to an attempt by a robot to deceive in a reciprocal game. The robot "bribed" participants by intentionally letting them win (i.e., changing behavior to the benefit of the participant) in one task, and afterwards asked them to help in a task

that would benefit the robot. Researchers reported that the robots' deceptive behavior had no effect on participants' behavior, as almost all participants agreed to help the robot regardless of its actions. Interestingly, participants rated the cheating robot as more likable than the honest one, perhaps, attributing its behavior to a prosocial intention (Sandoval et al., 2016).

Users like not only a "dishonest" robot, but also an erroneous robot, sometimes even more than they like one that does not make any mistakes. Mirnig et al. (2017) intentionally designed a robot that makes erroneous explanations, and compared users' liking and perceptions of anthropomorphism and intelligence. They found that the erroneous robot was liked more than a flawless one, and that other perceptions were not affected (Mirnig et al., 2017). Similarly, Ragni et al., (2016) found that people experienced more positive emotions toward a robot who demonstrated less than perfect memory skills in comparison to a flawless memorizing robot. While Ragni et al., (2016) suggested that higher liking could be explained by a lowered sense of competition with an erroneous robot, it is possible that uncanny valley theory also provides a valid explanation, suggesting that a human-like, flawless robot could induce higher levels of discomfort than one that makes mistakes (Groom et al., 2009). Future research should further explore the reasons for the positive emotional reaction toward imperfect functioning anthropomorphic robots.

Just as with anthropomorphic behavior, high-immediacy animal-like behaviors can also induce emotional trust. Examining human reactions to a dog-like robot, Lee, Park, and Song (2005) found that the robot's ability to improve its responsiveness had a significant effect on the robot's likability and humans' trust and increased the willingness to spend more time with it. It seems that behaviors that reflect social intelligence, such as social gestures, responsiveness, active listening, back-channeling, learning, and even cheating (for the benefit of the user) have more consistent and profound impact on emotional trust than the features related to the robot's appearance.

Furthermore, even when the initial trust is low, experiencing an interaction with a pro-social robot will increase trust.

Emotional trust in virtual AI

Tangibility

Research suggests that for virtual AI, tangibility has mostly positive effects on emotional trust (de Visser et al., 2016; Pak, Fink, Price, Bass, & Sturre, 2012; Qiu & Benbasat, 2009; Waytz et al., 2014). Chattaraman and Kwon (2014) found that the presence of a "persona" in a mock retail website significantly reduced the anxiety of older users, increasing perceived social support. Similarly, Qiu and Benbasat (2009) demonstrated that virtual embodiment of a recommendation agent significantly improved users' enjoyment and trust, increasing perceptions of social presence.

It seems that tangibility of virtual agents may even induce a physiological effect. De Visser et al. (2017) found that oxytocin had an impact on human trust in an anthropomorphic agent, leading participants to trust a virtual agent more than an embedded AI (i.e., AI that has no tangible identity). The connection of oxytocin to trust in the virtual agent suggests that people tend to perceive such agents as social actors, reacting similarly even on a physiological level.

Anthropomorphism

By contrast with tangibility, anthropomorphism has more mixed effects on emotional trust in virtual AI. Culley and Madhavan (2013) suggested that anthropomorphic characters are often depicted as capable of human qualities, including reasoning and motivation, which can induce very high expectations and initial trust. Since these expectations are unrealistic, however, high expectations of anthropomorphic characters are designed to fail. Ben Mimoun et al. (2012) suggested that poor calibration between virtual agents' appearance on commercial websites and

their actual performance drove customers' distrust and abandonment, which in turn caused the website owners to stop using the agents.

Anthropomorphic features provide the opportunity to manipulate virtual AI appearance in different ways, making it more attractive and thus increasing its likability (Bartneck et al., 2009; Beldad, De Jong, & Steehouder, 2010; Khan & Sutcliffe, 2014; Obaid, Salem, et al., 2016; Pak et al., 2012; Verberne, Ham, & Midden, 2015). For instance, Khan and Sutcliffe (2014) found that the visual representation of virtual AI has a significant impact on human compliance. By comparing agents presented through two slightly different female images, the authors found that the more visually-attractive agent was significantly more persuasive for both male and female participants. Personalization of AI interface features may also increase likability. Researchers have suggested that to be effective in the global market, a virtual agent should conform to different cultural preferences, including language, communication patterns, and facial characteristics such as those that are associated with an ethnicity (Culley & Madhavan, 2013). Some empirical studies support this notion, demonstrating that when the ethnic facial features of a virtual agent match a cultural group, it increases users' emotional trust (Obaid et al., 2016a). Relatedly, while examining the personalization of an agent's visual image, Verberne et al. (2015) found that when an agent was represented by a face whose features were adjusted based on the face of the user, the users reported higher levels of trust while using a driving simulation and were more willing to allow the agent to choose the route.

The issue of the anthropomorphism of a virtual agent leads to an additional question: to what extent does it matter if the virtual agent represents AI or another human? The interest in virtual platforms such as Second Life has made possible the use of avatars for representing humans in the virtual space in different contexts, including business interactions. Empirical research has

followed this trend, testing the effect of such representation on human reactions in general and trust in particular. In a meta-analysis of 32 studies, (Fox et al., 2015) found that when avatars were presented to users as humans, they were more influential than when they were presented as AI-based. Interestingly, the influence was more evident in objective (behavioral) rather than subjective (self-report) measures. It is important to note that the framing of the avatar as human or as AI had a stronger impact on participants' perceptions than the actual level of intelligence or control over the character (Von Der Pütten et al., 2010).

Looking further into the differences between human avatars and intelligent agents activated by AI, researchers have found that when talking to AI people tend to engage less in impression management and to disclose more sensitive personal information than when talking to a human (Gratch, Lucas, King, & Morency, 2014; Krämer et al., 2017; Lucas, Gratch, King, & Morency, 2014). This tendency implies the potential for higher emotional trust in AI than in humans.

Immediacy behaviors

The interactive abilities of virtual AI were mostly found to facilitate users' positive emotions, emotional trust and satisfaction. Kaptein and colleagues (2011) found that positive feedback about the ongoing conversation or social praise provided by an AI virtual agent increased its perceived friendliness. Dabholkar and Sheng (2012) demonstrated that an interactional recommendation agent made users get more involved in the process which led to higher satisfaction and trust. Matsui and Yamada (2016) revealed that a virtual agent that used hand gestures and expressive facial movements led to more positive emotional contagion than a virtual agent that was less expressive.

However, the effect of immediacy behaviors on emotional trust might be moderated by users' characteristics, such as need for social interaction. Ben Mimoun et al. (2017) found that for users with high need for interaction, the use of visually-present, interactive virtual AI (compared with a present but not interactive agent) had a positive effect on perceived system social presence and playfulness; however, for users with low need for social interaction, the use of virtual AI led to no effect on social presence and playfulness. Personalization of the immediacy behaviors to users' needs and preferences could be an important factor for increasing the positive effects of virtual AI.

An additional limitation to the effectiveness of immediacy behaviors was found by Groom et al. (2009) who showed that people experience more positive emotions toward agents whose behavior is not completely realistic. The researchers compared a human-like virtual agent with a pre-recorded human voice recommendation that demonstrated different variations of body and lip synchronized movements. They found that the agent was most liked when it engaged in only one of the synchronized behaviors (either lips or body movement), but not both. Researchers argued that these results are consistent with the Uncanny Valley theory (Mori, 1970), suggesting that when the looks and behavior of an artificial agent are too human-like, people experience discomfort and a sense of eeriness.

Emotional Trust in Embedded AI

Emotional trust in embedded AI can be built based on the reputation of the technology and the reputation of the organizations associated with it. For instance, Hengstler et al. (2016), in an analysis of eight case studies from the health and transportation industries, examined the way organizations aim to establish users' trust in AI. Using semi-structured interviews, the researchers concluded that firms promote trust in AI by connecting it to the reputation of the developing

organization and by making the technology more comprehensible, emphasizing its current and future usability and benefits. However, this study is limited to the perceived organizational intentions, without presenting the effectiveness of such tactics. Developers' reputation could be related to the perceived moral standards of the algorithm. Jago (2019) found that algorithmic decision was liked less than same decision made by a human due to its perceived lower authenticity and ethicality. Future studies should further examine the impact of moral perceptions and different persuasion tactics for individual users as well as for organizations that consider integrating AI into their processes and production.

Due to the low visibility of embedded AI, its impact on human emotions and emotion-related trust is less clear, and as of this writing there is very little empirical research that addresses how tangibility, anthropomorphism or immediacy impacts emotional trust in embedded AI. An example of an emotional reaction to revealing the use of AI could be found in the study on Facebook's news feed. Researchers reported that users that were unaware of algorithms being used felt surprise and anger (Eslami et al., 2015). After learning the features of the algorithm, the users became more active, gaining a higher level of control. This means that experienced negative emotions did not prevent the use of AI but lead to a more conscious utilization.

It is possible that the initial high trust that is evident from lab studies and the initial low trust that is evident from field studies (Alan et al., 2014; Linkov et al., 2017) are results of emotional reactions related to the perceived role of AI (i.e., assisting or threatening). However, more empirical research is needed to understand the antecedents of initial trust and the interplay between cognitive and emotional trust with regard to embedded AI.

General Discussion

Building Cognitive Trust in AI – Discussion and Future Research

Our review of the empirical research on cognitive trust in AI demonstrates that AI representation and the level of machine intelligence play an important role in the nature of the trust people develop. Examining the different dynamics of cognitive trust, our review reveals that for robotic AI, the trust trajectory is similar to that characterizing trust development in human relationships—it starts low and increases following direct, hands-on experience. However, for virtual and embedded AI, we see the opposite; most commonly, high initial trust drops as a result of experience (see Figure 1). The level of machine intelligence may moderate the development of trust, with a high level of intelligence leading to higher trust following use and experience. For robotic AI, a high level of machine intelligence generally leads to faster development of a high level of trust that can be resilient, even in cases of low reliability. For virtual and embedded AI, high machine intelligence offers the possibility of maintaining the initial high levels of trust (by meeting the high expectations) or minimizing the reduction of trust.

It is important to note that while the described trust trajectories capture the findings of the majority of studies to date, there are some studies that run contrary to these patterns. While some lab studies found high initial trust in robotic AI, some other studies, mostly field-based, found low levels of initial trust for virtual and embedded AI. A main concern with low levels of initial trust is its implication for disuse (refusal to use) or abuse (playing the algorithm) behaviors. Ironically, while disuse prevents the hands-on experience that has the potential to build trust, abuse of AI technology does provide such an opportunity. This means that when AI is highly intelligent, even abusive misuse may lead to establishing trust (while producing unintended behavior). Future research needs to take a long-term perspective in examining the effect of AI responsiveness to

different disuse and abuse behaviors on developing trust and the human-AI working relationship. In addition, there is a growing need for research in real-life settings, such as organizations that are already using AI in their management or decision-making systems. In lab studies, high initial trust could be facilitated by the controlled environment and the experimenter's involvement. Therefore, despite the valuable knowledge that is provided by lab studies, there is an urgent need to conduct more field studies where using AI is associated with greater personal risk for users.

Examining the overall effect of different functional characteristics on establishing cognitive trust in AI, we can conclude that for all AI representations the characteristics of the task the user and/or technology is performing play an important role; human trust is higher for issues that do not require social or emotional intelligence. This effect is consistent with the general assumptions regarding the advances of technology over humans and humans over technology (e.g., MABA-HABA). While perceptions of self-capabilities and expertise may moderate the beliefs regarding AI abilities to perform a specific task, future research should further examine the way immediacy behaviors influence these predispositions and judgments. The prosocial interactional behaviors of AI may increase the perceptions of its social intelligence, and thus increase the range of tasks for which AI-enabled technology could be perceived as an expert.

Transparency, which was rarely studied with regard to trust in robotic AI, was found to be highly important for establishing cognitive trust in virtual and embedded AI. Specifically, two types of transparency were found to be effective: explanations of how the algorithm works, and reflection of AI reliability. While understanding the ways in which AI makes decisions could be impossible under certain circumstances, the transparency of its level of ability and expected reliability may play a large role in the process of calibration between users' expectations and AI performance. Such transparency may lower the unrealistically high levels of initial trust that

sometimes form, which in turn may ease the actual use, by preventing the rapid drop in trust (also observed) and keeping the levels of trust stable during longer-term use. Future research should address more thoroughly the issue of transparency, especially in cases where the technology is still in earlier stages of development and could suffer from inconsistent performance.

In the past, reliability was considered to be the most important factor for adopting new technologies; however, the relationship between reliability and trust in AI could be complex, as low reliability does not always lead to low trust and disuse. On the one hand, the evidence for the importance of reliability for trust could be found in studies on embedded AI in which AI was a decision-making aid that was disused as a result of low reliability. However, it is important to note that in these lab studies, AI was presented similarly to less-complex technologies, had no agency or physical representation, and was not engaged in any immediacy behaviors, or exhibiting any other signs of high machine intelligence. On the other hand, for robotic AI, high machine intelligence and immediacy behaviors were found to moderate the effect of reliability on trust, with immediacy behaviors increasing trust despite the low reliability demonstrated by erroneous actions. Furthermore, erroneous actions may even increase emotional trust and liking (considered more below), especially for robotic and virtual AI, raising important questions regarding the potential for manipulating human trust. Thus, future research should further examine the relationship between AI reliability and trust across different AI representations, considering machine intelligence as a possible moderator, and examining additional moderators, such as the consistency of performance levels, expectations and prior beliefs, and perceived level of risk by the user.

The positive role of immediacy behaviors has been examined most in research on robotic AI, as these were perceived as indicators of high machine intelligence. These behaviors appear to

be sufficient for establishing trust (even in cases when robotic AI acts erroneously); however, the psychological mechanism that explains these relationships is not clear. It is possible that people perceived the immediacy behaviors as a sign of the AI agent's level of intelligence, or as a recognition of their own value. Similar to the notion famously expressed by Theodore Roosevelt, "Nobody cares how much you know, until they know how much you care," it appears that users are highly responsive to indications that the technology "cares" about and is responding to them. Future research should explore human biases related to various aspects of immediacy behaviors, such as listening and personalization, not only in order to understand its functionality for human trust, but also to prevent unethical use of such tactics.

Building Emotional Trust in AI - Discussion and Future Research

This review provides several important insights regarding the factors that influence emotional trust across different AI representations. Focusing on appearance and behavioral factors, the existing literature examines emotional trust mostly with regards to robotic and virtual AI, with very few studies on emotional trust in embedded AI. However, there is a growing need to examine the aspects that may influence emotional trust in embedded AI, especially when users had no prior awareness of an AI presence (Powers, 2017). Influential features could include the form and timing that governs the AI presence and explanation, or the relative role of AI developers' reputation. Even when emotional trust is not expected to be the main factor driving AI use, it could significantly facilitate or moderate the effect of cognitive factors. Therefore, future research should examine the direct and moderating roles of emotional trust in embedded AI.

With respect to virtual and robotic representations, it is interesting to note that the effect of tangibility on emotional trust significantly varies. For virtual AI, tangibility was found to have a mostly positive effect, facilitating social presence, increasing liking, positive feelings and

emotional trust. By contrast, for robotic AI, it's physical presence may also induce negative emotions, such as fear. Future research should further explore the importance of the negative feelings evoked by AI tangibility, testing their possible implications for disuse or abuse, as well as for long-term relationships. It is possible that while decreasing initial trust, the impact of tangibility for developing trust over a longer period is insignificant.

Interestingly, the effect of anthropomorphism also differs across representations in terms of implications for emotional trust. For virtual AI, anthropomorphism was found to have a positive impact. For robotic AI, the evidence is mixed, finding both positive and negative effects of anthropomorphism on emotional trust. On the one hand, people like anthropomorphic robots more than mechanical-looking robots, but on the other hand, anthropomorphic robots may evoke negative feelings, discomfort and a sense of eeriness. It seems that at least part of this negative impact could be explained by a mismatch between human-like appearance and low machine intelligence. However, it is also possible that human-like appearance matched with perfect performance will also lead to negative emotional reactions, as suggested by the Uncanny Valley theory. The greater likability of a mistake-making robot compared to a highly reliable robot suggests that high intelligence and perfect performance may intimidate users, leading to feelings of discomfort and distrust. A similar effect was found for virtual AI, demonstrating that users prefer the imperfect agent that does not demonstrate a full match between voice and movement. Future research must further examine the boundaries for the positive and negative effects of the match between anthropomorphic appearance and interactive immediacy behaviors for emotional trust.

Cognitive versus Emotional Trust in AI

The reviewed research indicates that factors that influence emotional trust differ from those influencing cognitive trust, and some factors may even have different implications for cognitive and emotional trust. Transparency and reliability, while having some effect on cognitive trust in AI, remain relatively unexplored with regards to emotional trust. Furthermore, when studied, in some cases lower reliability was found to have the opposite effect to what was expected, exhibiting a positive impact on emotional trust in robotic AI.

The effect of tangibility mostly plays a positive role for cognitive trust; however, it is not so for emotional trust, as physical presence could be perceived as threatening, and virtual presence may evoke unrealistic expectations. High anthropomorphism was found to generate perceptions of high machine intelligence; however, it was mostly studied in the context of emotional trust, and might also have negative emotional implications, especially in the case of a mismatch between representation and AI capabilities. The impact of immediacy behaviors is also different for cognitive and emotional trust. For cognitive trust, immediacy behaviors were always shown to have a positive effect, while for emotional trust some immediacy behaviors had the opposite effect, making people uncomfortable, such as when there was too close of a match between human-like appearance and behaviors.

It seems that emotional and cognitive trust differ even with regard to the impact of task characteristics. With respect to receiving advice, users have more cognitive trust in AI than in human advice on technical issues, but not on issues requiring social knowledge. However, when engaging in self-disclosure, users tend to be more honest and open with AI than with other humans, disclosing more personal information and engaging less in impression management behaviors.

Studies on emotional trust have more extensively explored the impact of users' predispositions toward technology than have studies on cognitive trust. Thus, with regard to emotional trust, we know that the need for social interaction and NARS (negative attitude toward robots) both have important implications for trust development. In contrast, studies on cognitive trust mentioned users' expectations for level of performance, especially for virtual and embedded AI, yet rarely addressed its source nor any other stable individual differences that might affect user responses to AI technology.

Taken together, these discrepancies clearly indicate the differences between cognitive and emotional trust development in AI, leading to the question of their relative impact on the actual use of AI. However, the extant research findings on whether emotional trust factors, such as likability, are more or less important than cognitive trust factors are mixed. Matsui and Yamada (2019) tested the impact of virtual agent knowledge and social gestures on participants' emotions, perceptions and trust, and found that an increase in participants' positive affect facilitated their trust even when the perception of agent intelligence was low (suggesting more relative importance for emotional over cognitive trust features). In contrast, Wang et al. (2016) investigated cognitive and emotional trust and found that perceived professionalism was important for establishing emotional trust. Future studies should address the emotional and cognitive aspects together, suggesting the conditions under which there are synergies, or when emotional factors will be more important than cognitive and vice versa.

Researchers seeking to understand emotional and cognitive trust in AI may benefit from taking into consideration human-human swift trust and the first impression literature. This research suggests that our impressions are driven by the perceptions of the counterpart as having good intentions (warmth) and the ability to pursue the intentions (competence; Cuddy, Glick, &

Beninger, 2011). Trust in AI is likely to depend on both AI's likability and its *perceived intelligence*. However, in contrast to human trustees, for AI, the features influencing these perceptions could be easily manipulated. While for prior generations of technology the role of the interface was mostly aimed at improving the ease of use, the perceived warmth/likability of AI's representation has tremendous potential to influence human emotions (e.g., Krämer et al., 2017; Looije et al., 2010). An AI agent's attractiveness and its visual similarity to a user (Khan & Sutcliffe, 2014; Verberne et al., 2015), as well as the features of a human-like robot's face (Zhang et al., 2010) may evoke unconscious emotional reactions, similar to perceptions of warmth, driving perceptions of AI's benevolence and positive intentions, and in this way could influence human trust and behavior. These reactions, being largely unrelated to the actual AI capabilities and intent, could create impressions that are difficult to change and require mindful ethical consideration and further research.

The reviewed research also suggests some new paths for further understanding trust among humans. It could be useful to consider the different cognitive and emotional implications of human physical presence, conceptualizing presence as a continuum rather than as a category (i.e., co-located vs. distributed), and further exploring the moderating factors in the tangibility-trust relationship. The relative impact of physical presence has already drawn some attention from researchers studying group diversity and virtual communication; however, addressing the presence as a continuum and focusing on moderating factors could increase understanding of the nuances and underlying psychological mechanisms that explain tangibility-trust relations. In addition, the notion of calibration between early expectations, shaping initial trust, and actual capability and performance could be useful for encouraging the development of trust, especially for individuals and groups working at a geographical distance. Finally, as machine intelligence increases, and we

develop a better understanding of AI's possible interference in human-human relations, it would be possible to use AI for facilitating positive initial trust among humans by encouraging safe and mutually acceptable conversational content (Isbister, Nakanishi, Ishida, & Nass, 2000), and increasing mutual trust by facilitating more fair offers in negotiations and decision-making (de Melo, Marsella, & Gratch, 2016, 2017; M. K. Lee & Baykal, 2017).

Integrative framework for current and future research on AI

As demonstrated in this review, empirical research on trust in AI is distributed across different fields and AI is represented to users in different forms. The various types of representation could be considered as the material presence associated with AI, while the tangibility by itself has a dramatic effect on AI trust and acceptance (e.g., de Visser et al., 2016). Following the reviewed empirical studies and based on the theoretical developments discussed (K. M. Lee, Jung, et al., 2006), we propose an AI embodiment framework that can help guide the integration of multidisciplinary knowledge on human-AI relations and facilitate future research. While we used this framework in a categorical manner while reviewing the literature, we suggest that since it reflects the material representation of AI, it could be seen as a continuum that starts with a complete physical presence (robotic representation), gradually diminishes to a virtual presence such as a 2D agent, image, voice or text, and ends with an absence of any distinguished AI presence (embedded representation). Addressing AI representations as a continuum allows researchers to examine cases in which the representation does not clearly belong to just one of the categories; for example, cases in which an AI agent has both physical and virtual representation, or when the identification of AI is clear to some people but not to others. These marginal cases are important as they can assist in better understanding the way representation and tangibility influence trust and human behavior.

Our framework also suggests that the level of machine intelligence plays an important moderating role for human trust. In addition to enabling immediacy behaviors, such as responsiveness, high machine intelligence is required for complex functions and higher control over tasks. It may moderate the way trust is being established as well as the importance of different dimensions for establishing trust (such as reliability).

As reflected by the reviewed research, the level and type of embodiment, combined with the level of machine intelligence encompassed in the technology, have significant implications for the users' perceptions and feelings that lead to the type and level of trust that users develop. We propose that this framework could be useful to researchers for further understanding not only trust, but also additional factors such as cooperation and reliance in human-AI relations.

Additional directions for future research

Multidisciplinary research on AI has developed dramatically during the last twenty years, moving from a strict focus on technological objectives toward an interest in the human users' perspective. Whereas in the past, AI researchers considered human cognition and behavior only for the purposes of developing mathematical models that would allow AI to mimic human logic, the relative maturity of technologies has led to a shift toward a human-centered approach (Jaimes, Gatica-Perez, Sebe, & Huang, 2007) that considers the needs, perceptions and behaviors of a human user in the design. This approach was driven partially by potential users' difficulty in trusting AI technology and by the willingness of AI developers to address these issues in order to increase collaboration with AI (Gross, 2010; Sierhuis et al., 2003). AI-enabled technology presents an unprecedented opportunity for technology to develop a responsive, adaptive, supportive "relationship" with human users that could yield a wealth of benefits but also be a source of significant threat.

The human-centered approach creates a great opportunity for collaboration among researchers interested in the evolution and integration of new technologies from different disciplinary perspectives. Although most of the current research on human-AI dynamics is being conducted by cognitive engineering and information systems researchers, organizational research may contribute an important perspective that would allow consideration of micro- and macro-level factors as well as the short- and long-term processes affected by the introduction of AI into organizations. For example, how does the implementation of AI-guided hiring and evaluation change the relationships of workers with their jobs? With their co-workers? With their supervisors? How does it shape the distribution of power in the organization?

Despite the fact that the field of organizational behavior and management usually focuses on well-established phenomena rather than on unfolding events, the opportunity to make a significant impact on the way AI is currently developed suggests that we use the aggregated knowledge to provide theoretical models relevant to the future of organizations. Reviewing the existing empirical research, we aimed to provide organization and management researchers some needed perspective to join the emerging multidisciplinary discussion that may determine the way organizations will integrate and use AI in the future.

An additional contribution that could be made by fostering interdisciplinary collaboration on these topics is an improvement in research methodology, including proper study design with human subjects and advanced statistical analyses. Most importantly, there is an urgent need for addressing the great variance in measures used to assess human trust in AI. While many researchers invested in developing new scales and behavioral measures for trust (Bartneck et al., 2009; Charalambous, Fletcher, & Webb, 2016; Headleand et al., 2016; Ho & MacDorman, 2010; Miller et al., 2016; Ullman & Malle, 2018; Walker, Verwey, & Martens, 2018), the lack of open dialogue

across different disciplines regarding the issue of measurement might discourage researchers from collaborating and limit research implications to a specific discipline.

This review focused on the trust of an individual user in AI, which has been the major focus of the existing literature in this area. However, as some researchers have noted (e.g., Yagoda & Gillan, 2012), the development of trust in AI is often not limited to an individual user, but also involves relationships with other humans and machines being directly or indirectly influenced by AI. Examining AI as part of a complex system, such as a team or a network, would allow researchers to address the way people establish true relationships with AI within organizations, as well as the way AI may change the relationship between humans, and between humans and other machines. Furthermore, as AI behavior is not deterministic, scholars need to examine the way it changes based on human-AI interactions (Rahwan et al., 2019), facilitating knowledge on the evolving relationships. Future research should take into consideration a multi-layer perspective of trust in AI that could more accurately explain human behavior.

It is important to note that, so far, many of the analyses examining the future of AI integration in organizations take the technological perspective, focusing on the current maturity of specific technology that is rapidly improving and changing. From this perspective, the smarter the AI, the smarter the organization can be. A human-centered approach needs to consider AI-integration from the employees' perspective, taking into consideration elements that facilitate human trust, and the meaningfulness and importance of a specific task to the employees. Past research demonstrates that the adjustment of employees to new technologies is a key factor in translating technological advances into business revenue (Davenport & Short, 1990). AI is unlikely to be different. Therefore, when making decisions regarding the tasks to be outsourced to AI, managers should consider not only the available technological capabilities, but also the human

participants, their interests and incentives, and the ways to gain their trust and improve their productivity via collaboration with AI.

An additional aspect that should be considered by organizational researchers relates to the skills and characteristics required by future organizational leaders, who will manage not only human employees, but also complex systems of different algorithms collaborating among themselves and with humans. Guiding the new generation of leaders who must be technologically-educated, there is a strong need to make sure that skills to manage human employees will keep playing an important role in business programs. Keeping the "human in the loop" is an essential part of AI integration; therefore, future leaders should be able to manage machine-machine, human-human, *and* human-machine teams.

This review presents recent multidisciplinary empirical research on cognitive and emotional trust in AI in its various representations. By presenting the main findings, providing a research framework and highlighting the most promising future directions we aim to encourage researchers to explore the various aspects of human-AI interaction in order to facilitate a human-centered, ethical and safe integration of AI within organizations.

Acknowledgements

The work on this article was sponsored by the Defense Advanced Research Projects Agency and the Army Research Office and was accomplished under Grant Number W911NF-17-1-0104 and W911NF-20-1-0006. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, the Army Research Office, or the

References

Alan, A., Costanza, E., Fischer, J., Ramchurn, S. D., Rodden, T., & Jennings, N. R. (2014). A field study of human-agent interaction for electricity tariff switching. *13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, 965–972. Retrieved from www.ifaamas.org

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Andrews, P. Y. (2012). System personality and persuasion in human-computer dialogue. *ACM Transactions on Interactive Intelligent Systems*, *2*(2), 1–27. https://doi.org/10.1145/2209310.2209315

Andrist, S., Bohus, D., Yu, Z., & Horvitz, E. (2016). Are you messing with me? Querying about the sincerity of interactions in the open world. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*(1), 409–410. https://doi.org/10.1109/HRI.2016.7451780

Appel, M., Weber, S., Krause, S., & Mara, M. (2016). On the eeriness of service robots with emotional capabilities. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 411–412. https://doi.org/10.1109/HRI.2016.7451781

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*(1), 41–52. https://doi.org/10.1007/s12369-010-0082-7

Baraglia, J., Cakmak, M., Nagai, Y., Rao, R., & Asada, M. (2016). Initiative in robot assistance during collaborative task execution. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 67–74. https://doi.org/10.1109/HRI.2016.7451735

Bartneck, C., Kulić, D., Croft, E., & 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. https://doi.org/10.1007/s12369-008-0001-3

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2006). The influence of people's culture and prior experiences with Aibo on their attitude towards robots. *AI & SOCIETY*, *21*(1–2), 217–230. https://doi.org/10.1007/s00146-006-0052-7

Beldad, A., De Jong, M., & Steehouder, M. (2010). How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. *Computers in Human Behavior*, *26*(5), 857–869. https://doi.org/10.1016/j.chb.2010.03.013

Ben Mimoun, M. S., Poncin, I., & Garnier, M. (2012). Case study—Embodied virtual agents: An analysis on reasons for failure. *Journal of Retailing and Consumer Services*, *19*(6), 605–612. https://doi.org/10.1016/J.JRETCONSER.2012.07.006

Ben Mimoun, M. S., Poncin, I., & Garnier, M. (2017). Animated conversational agents and e-consumer productivity: The roles of agents and individual characteristics. *Information and Management*, *54*(5), 545–559. https://doi.org/10.1016/j.im.2016.11.008

Bickmore, T., Pfeifer, L., & Schulman, D. (2011). *Relational Agents Improve Engagement and Learning in Science Museum Visitors*. https://doi.org/10.1007/978-3-642-23974-8_7

Bickmore, T. W., Vardoulakis, L. M. P., & Schulman, D. (2013). Tinker: A relational agent museum guide. *Autonomous Agents and Multi-Agent Systems*, *27*(2), 254–276. https://doi.org/10.1007/s10458-012-9216-7

Birnbaum, G. E., Mizrahi, M., Hoffman, G., Reis, H. T., Finkel, E. J., & Sass, O. (2016). Machines as a source of consolation: Robot responsiveness increases human approach behavior and desire for companionship. *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, 165–171. Retrieved from https://dl.acm.org/citation.cfm?id=2906861

Bradshaw, J. M., Feltovich, P., & Johnson, M. (2011). Human-Agent Interaction. *Handbook of HumanMachine Interaction*, 293–302. Retrieved from http://books.google.com/books?hl=en&lr=&id=4opHlu05SNIC&oi=fnd&pg=PA283&dq=Human-agent+interaction&ots=vxrpDdLbSa&sig=07dujtzGjIcBLlZ6FVH33HjrWos

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534.

Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, 43–47. https://doi.org/10.1257/pandp.20181019

Burr, C., Cristianini, N., & Ladyman, J. (2018). An Analysis of the Interaction Between Intelligent Software Agents and Human Users. In *Minds and Machines*. https://doi.org/10.1007/s11023-018-9479-0

Camilleri, A. R., Cam, M.-A., & Hoffmann, R. (2007). Nudges and signposts - The effect of smart defaults and pictographic risk inofrmation on retirement saving investment choices. *SRNN, 111*(2), 154–162.

Carlson, Z., Sweet, T., Rhizor, J., Poston, J., Lucas, H., & Feil-seifer, D. (2015). Team-building activities for heterogeneous groups of humans and robots. In A. Tapus, E. André, J. Martin, F. Ferland, & M. Ammi (Eds.), *Social Robotics. ICSR 2015. Lecture Notes in Computer Science, vol 9388* (pp. 113–123). Springer, Cham.

Chao, C. Y., Chang, T. C., Wu, H. C., Lin, Y. S., & Chen, P. C. (2016). The interrelationship between intelligent agents' characteristics and users' intention in a search engine by making beliefs and perceived risks mediators. *Computers in Human Behavior*, *64*, 117–125. https://doi.org/10.1016/j.chb.2016.06.031

Charalambous, G., Fletcher, S., & Webb, P. (2016). The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration. *International Journal of Social Robotics*, *8*(2), 193–209. https://doi.org/10.1007/s12369-015-0333-8

Chattaraman, V., Kwon, W.-S., E. Gilbert, J., & Li, Y. (2014). Virtual shopping agents. *Journal of Research in Interactive Marketing*, *8*(2), 144–162. https://doi.org/10.1108/JRIM-08-2013-0054

Chen, J. Y. C., & Barnes, M. J. (2014). Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems, 44*(1), 13–28.

Cho, J. E., & Hu, H. (2009). The effect of service quality on trust and commitment varying across generations. *International Journal of Consumer Studies*, *33*(4), 468–476. https://doi.org/10.1111/j.1470-6431.2009.00777.x

Cormier, D., Young, J., Nakane, M., Newman, G., & Durocher, S. (2013). Would you do as a robot commands? An obedience study for human-robot interaction. *International Conference on Human-Agent Interaction*, I-3–1.

Cotter, K., Cho, J., & Rader, E. (2017). Explaining the news feed algorithm. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*, 1553–1560. https://doi.org/10.1145/3027063.3053114

Crisp, C. B., & Jarvenpaa, S. L. (2013). Swift trust in global virtual teams: Trusting beliefs and normative actions. *Journal of Personnel Psychology*, *12*(1), 45–56. https://doi.org/10.1027/1866-5888/a000075

Cuddy, A. J. C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, *31*, 73–98. https://doi.org/10.1016/j.riob.2011.10.004

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior*, *29*(3), 577–579. https://doi.org/10.1016/J.CHB.2012.11.023

Dabholkar, P. A., & Sheng, X. (2012). Consumer participation in using online recommendation agents: effects on satisfaction, trust, and purchase intentions. *The Service Industries Journal*, *32*(9), 1433–1449. https://doi.org/10.1080/02642069.2011.624596

Danaher, J. (2017). Will life be worth living in a world without work? Technological ynemployment and the meaning of life. *Science and Engineering Ethics*, *23*(1), 41–64. https://doi.org/10.1007/s11948-016-9770-5

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *International Joint Conference on Artificial Intelligence*, (January), 4691–4697. https://doi.org/10.1111/j.1365-2796.2007.01905.x

Davenport, Thomas, H., & Short, James, E. (1990). The new industrial engineering: Information technology and business process redesign. *Sloan Management Review*, *31*(4), 11–27. Retrieved from https://search.proquest.com/docview/224963315?pq-origsite=gscholar

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003. https://doi.org/10.1287/mnsc.35.8.982

Davis, G. F. (2019). How to communicate large-scale social challenges: The problem of the disappearing American corporation. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(16), 7698–7702. https://doi.org/10.1073/pnas.1805867115

de Melo, C. M., Marsella, S., & Gratch, J. (2016). "Do as I say, not as I do:" Challenges in delegating decisions to automated agents. *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, (Aamas), 949–956.

de Melo, C. M., Marsella, S., & Gratch, J. (2017). Increasing fairness by delegating decisions to autonomous agents. *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, (Aamas), 419–425. Retrieved from http://dl.acm.org/citation.cfm?id=3091188

de Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R., … Krueger, F. (2017). A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Human Factors*, *59*(1), 116–133. https://doi.org/10.1177/0018720816687205

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

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

de Winter, J. C. F., & Dodou, D. (2014). Why the Fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, *16*(1), 1–11. https://doi.org/10.1007/s10111-011-0188-1

Demir, M., McNeese, N. J., & Cooke, N. J. (2017). Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research*, *46*, 3–12. https://doi.org/10.1016/J.COGSYS.2016.11.003

Demir, M., McNeese, N. J., Cooke, N. J., Ball, J. T., Myers, C., & Frieman, M. (2015). Synthetic teammate communication and coordination with humans. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 951–955. https://doi.org/10.1177/1541931215591275

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot

failures and feedback on real-time trust. *ACM/IEEE International Conference on Human-Robot Interaction*, 251–258. https://doi.org/10.1109/HRI.2013.6483596

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., … Yanco, H. (2012). Effects of changing reliability on trust of robot systems. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '12*, 73. https://doi.org/10.1145/2157689.2157702

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, mnsc.2016.2643. https://doi.org/10.1287/mnsc.2016.2643

Duarte, J., Siegel, S., & Young, L. (2012). Trust and Credit: The Role of Appearance in Peer-to-peer Lending. *Review of Financial Studies*, *25*(8), 2455–2484. https://doi.org/10.1093/rfs/hhs071

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

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., … Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 153–162.

https://doi.org/10.1145/2702123.2702556

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). The influence of agent reliability on trust in human-agent collaboration. *ACM International Conference Proceeding Series*, *369*. https://doi.org/10.1145/1473018.1473028

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, *28*(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Fenster, M., Zuckerman, I., & Kraus, S. (2012). Guiding user choice during discussion by silence, examples and justifications. *Frontiers in Artificial Intelligence and Applications*, *242*(January), 330–335. https://doi.org/10.3233/978-1-61499-098-7-330

Ferràs-Hernández, X. (2018). The future of management in a world of electronic brains. *Journal of Management Inquiry*, *27*(2), 260–263. https://doi.org/10.1177/1056492617724973

Fox, J., Ahn, S. J. (Grace), Janssen, J. H., Yeykelis, L., Segovia, K. Y., & Bailenson, J. N. (2015). Avatars versus agents: A meta-analysis quantifying the effect of agency on social influence, human-computer interaction. *Human-Computer Interaction*, *30*(5), 401–432. https://doi.org/10.1080/07370024.2014.921494

Frantz, R. (2003). Herbert Simon. Artificial intelligence as a framework for understanding intuition. *Journal of Economic Psychology*, *24*(2), 265–277. https://doi.org/10.1016/S0167-4870(02)00207-6

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. *2007 International Symposium on Collaborative Technologies and*

Systems, 106–114. https://doi.org/10.1109/CTS.2007.4621745

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., & Ivaldi, S. (2016). Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Computers in Human Behavior*, *61*, 633–655. https://doi.org/10.1016/J.CHB.2016.03.057

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. *Cognition Technology and Work*, *14*(39), 49. https://doi.org/10.1007/s10111-011-0194-3

Glass, A., McGuinness, D. L., & Wolverton, M. (2008). Toward establishing trust in adaptive agents. *Proceedings of the 13th International Conference on Intelligent User Interfaces - IUI '08*, 227. https://doi.org/10.1145/1378773.1378804

Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F., & Shah, J. A. (2015). Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots*, *39*(3), 293–312. https://doi.org/10.1007/s10514-015-9457-9

Graetz, G., & Michaels, G. (2018). *The Review of Economics and Statistics ROBOTS AT WORK*. https://doi.org/10.1162/rest_a_00754

Gratch, J., Lucas, G., King, A., & Morency, L.-P. (2014). It's only a computer: The impact of human-agent interaction in clinical interviews. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, (AAMAS), 85–92.

Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J. K., & Robles, E. (2009). Evaluating the effects of behavioral realism in embodied agents. *International Journal of Human*

Computer Studies, *67*(10), 842–849. https://doi.org/10.1016/j.ijhcs.2009.07.001

Gross, T. (2010). Towards a new human-centred computing methodology for cooperative ambient intelligence. *Journal of Ambient Intelligence and Humanized Computing*, *1*(1), 31–42. https://doi.org/10.1007/s12652-009-0004-4

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Haring, K. S., Silvera-Tawil, D., Watanabe, K., & Velonaki, M. (2016). The influence of robot appearance and interactive ability in HRI: A cross-cultural study. *Lecture Notes in Computer Science*, *9979 LNAI*, 392–401. https://doi.org/10.1007/978-3-319-47437-3_38

Headleand, C. J., Jackson, J., Williams, B., Priday, L., Teahan, W. J., & Ap Cenydd, L. (2016). How the perceived identity of a NPC companion influences player behavior. *Lecture Notes in Computer Science*, *9590*, 88–107. https://doi.org/10.1007/978-3-662-53090-0_5

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI, 13)*, 210–217. Eindhoven, The Netherlands.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105–120. https://doi.org/10.1016/J.TECHFORE.2015.12.014

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose Job Is It Anyway? A Study of Human-

Robot Interaction in a Collaborative Task. *Human-Computer Interaction*, *19*, 151–181. Retrieved from https://pdfs.semanticscholar.org/6bcd/1edc188d10f5981884752a22965cbddd9cf8.pdf?_ga=2.104114819.2064917875.1561849320-1152773079.1561513581

Ho, C.-C., & 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. https://doi.org/10.1016/J.CHB.2010.05.015

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffman, G., & Breazeal, C. (2007). Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. *Proceeding of the ACM/IEEE International Conference on Human-Robot Interaction - HRI '07*, 1. https://doi.org/10.1145/1228716.1228718

Hollis, V., Pekurovsky, A., Wu, E., & Whittaker, S. (2018). On being told how we feel: How algorithmic sensor feedback influences emotion perception. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–31. https://doi.org/10.1145/3264924

Huang, M. H., & Rust, R. T. (2018). Artificial Intelligence in service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Isbister, K., Nakanishi, H., Ishida, T., & Nass, C. (2000). Helper agent. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '00*, (April 2000), 57–64.

https://doi.org/10.1145/332040.332407

Jacq, A., Lemaignan, S., Garcia, F., Dillenbourg, P., & Paiva, A. (2016). Building successful long child-robot interactions in a learning context. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 239–246. https://doi.org/10.1109/HRI.2016.7451758

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries*, *5*(1), 38–56. https://doi.org/10.5465/amd.2017.0002

Jaimes, A., Gatica-Perez, D., Sebe, N., & Huang, T. S. (2007). Guest editors' introduction: Human-centered computing - toward a human revolution. *Computer*, *40*(5), 30–34. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/MC.2007.169

Jiang, W., Fischer, J. E., Greenhalgh, C., Ramchurn, S. D., Wu, F., Jennings, N. R., & Rodden, T. (2014). Social implications of agent-based planning support for human teams. *2014 International Conference on Collaboration Technologies and Systems, CTS 2014*, 310–317. https://doi.org/10.1109/CTS.2014.6867582

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *The Academy of Management Review*, *23*(3), 531–546. Retrieved from https://www.jstor.org/stable/pdf/259293.pdf

Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., & Breazeal, C. (2013). Engaging Robots: Easing Complex Human-Robot Teamwork using Backchanneling. *Computer-Supported Cooperative Work*, 1555–1566. https://doi.org/10.1145/2441776.2441954

Kaplan, J. (2015). *Humans need not apply : A guide to wealth and work in the age of artificial*

intelligence. Yale University Press.

Kaptein, M., Markopoulos, P., de Ruyter, B., & Aarts, E. (2011). Two acts of social intelligence: The effects of mimicry and social praise on the evaluation of an artificial agent. *AI and Society*, *26*(3), 261–273. https://doi.org/10.1007/s00146-010-0304-4

Kellogg, K., Valentine, M., & Christin, A. (2019). Algorithms At Work: the New Contested Terrain of Control. *Academy of Management Annals*. https://doi.org/10.5465/annals.2018.0174

Khan, R. F., & Sutcliffe, A. (2014). Attractive agents are more persuasive. *International Journal of Human-Computer Interaction*, *30*(2), 142–150. https://doi.org/10.1080/10447318.2013.839904

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *CHI*. https://doi.org/10.1145/2858036.2858402

Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *Management Information Systems Quarterly*, *30*(4), 941–960. Retrieved from https://www.jstor.org/stable/pdf/25148760.pdf

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A conversational agent as museum guide - Design and evaluation of a real-world application. *Lecture Notes in Computer Science*, *3661 LNAI*, 329–343. https://doi.org/10.1007/11550617_28

Krämer, N. C., Lucas, G., Schmitt, L., & Gratch, J. (2017). Social snacking with a virtual agent – On the interrelation of need to belong and effects of social responsiveness when interacting with artificial entities. *International Journal of Human Computer Studies*, *109*, 112–121.

https://doi.org/10.1016/j.ijhcs.2017.09.001

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies*, *64*(10), 962–973. https://doi.org/10.1016/J.IJHCS.2006.05.002

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. https://doi.org/10.1111/j.1460-2466.2006.00318.x

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 205395171875668. https://doi.org/10.1177/2053951718756684

Lee, M. K., & Baykal, S. (2017). Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. *CSCW '17*. https://doi.org/10.1145/2998181.2998230

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. *CHI 2015*. https://doi.org/10.1145/2702123.2702548

Leonardi, Paul, M. (2009). Why do people reject new technologies and stymie organizational changes of which they are in favor? *Human Communication Research*, *35*(3), 407–441. https://doi.org/10.1111/j.1468-2958.2009.01357.x

Lewandowsky, S., Mundy, M., & Tan, G. P. A. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, *6*(2), 104–123.

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23–37. https://doi.org/10.1016/J.IJHCS.2015.01.001

Linkov, F., Sanei-Moghaddam, A., Edwards, R. P., Lounder, P. J., Ismail, N., Goughnour, S. L., … Comerci, J. T. (2017). Implementation of hysterectomy pathway: Impact on complications. *Women's Health Issues*, *27*(4), 493–498. https://doi.org/10.1016/J.WHI.2017.02.004

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *Journal of Strategic Information Systems*, *24*(3), 149–157. https://doi.org/10.1016/j.jsis.2015.08.002

Logg, J. M., Minson, J. A., & Moore, D. A. (2018). *Algorithm appreciation: People prefer algorithmic to human judgment* (No. 17–086). Retrieved from https://www.hbs.edu/faculty/Publication Files/17-086_610956b6-7d91-4337-90cc-5bb5245316a8.pdf

Looije, R., Neerincx, M. A., & Cnossen, F. (2010). Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies*, *68*(6), 386–397. https://doi.org/10.1016/J.IJHCS.2009.08.007

Lucas, G. M., Gratch, J., King, A., & Morency, L. P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 125–132. https://doi.org/10.1109/HRI.2016.7451743

Manzey, D., Reichenbach, J., & 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. https://doi.org/10.1177/1555343411433844

Martelaro, N., Jung, M., & Hinds, P. (2015). Using Robots to Moderate Team Conflict. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts - HRI'15 Extended Abstracts*, 271–271. https://doi.org/10.1145/2701973.2702094

Matsui, T., & Yamada, S. (2019). Designing Trustworthy Product Recommendation Virtual Agents Operating Positive Emotion and Having Copious Amount of Knowledge. *Frontiers in Psychology*, *10*, 675. https://doi.org/10.3389/fpsyg.2019.00675

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an

effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114

Mayer, R. C., Davis, J. H., & Schoorman, D. F. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, *38*(1), 24–59. Retrieved from https://journals.aom.org/doi/pdf/10.5465/256727

McCarthy, J., & Feigenbaum, E. A. (1990). In Memoriam: Arthur Samuel: Pioneer in Machine Learning. *AI Magazine*, *11*(3), 10–10. https://doi.org/10.1609/AIMAG.V11I3.840

Mehrabian, A. (1967). Attitudes inferred from non-immediacy of verbal communications. *Journal of Verbal Learning and Verbal Behavior*, *6*(2), 294–295. https://doi.org/10.1016/S0022-5371(67)80113-0

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral measurement of trust in automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1849–1853. https://doi.org/10.1177/1541931213601422

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers Robotics AI*, *4*(MAY). https://doi.org/10.3389/frobt.2017.00021

Möhlmann, M., & Zalmanson, L. (2017). Hands on the Wheel: Navigating Algorithmic Management and Uber Drivers ' Autonomy. *Proceedings of the Thirty Eighth International*

Conference on Information Systems (ICIS 2017), (December), 1–17. Seoul, South Korea.

Moran, S., Pantidi, N., Bachour, K., Fischer, J. E., Flintham, M., Rodden, T., … Johnson, S. (2013). Team reactions to voiced agent instructions in a pervasive game. *Proceedings of the 2013 International Conference on Intelligent User Interfaces - IUI '13*. https://doi.org/10.1145/2449396.2449445

Mori. (1970). Bukimi no tani [The uncanny valley]. *Energy*, *7*(4), 33–35.

Mumm, J., & Mutlu, B. (2011). Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Computers in Human Behavior*, *27*(5), 1643–1650. https://doi.org/10.1016/j.chb.2011.02.002

Murray, A., Rhymer, J., & Sirmon, D. (2019). Humans and agentic technologies: Toward a theory of conjoined agency in organizational routines. In ACM (Ed.), *Collective Intelligence*. Pittsburgh, PA.

Ng, K.-Y., & Chua, R. Y. J. (2006). Do I contribute more when I trust more? Differential effects of cognition- and affect-based trust. *Management and Organization Review*, *2*(01), 43–66. https://doi.org/10.1111/j.1740-8784.2006.00028.x

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies*, *7*(3), 437–454. https://doi.org/10.1075/is.7.3.14nom

Obaid, M., Salem, M., Ziadee, M., Boukaram, H., Moltchanova, E., & Sakr, M. (2016). Investigating effects of professional status and ethnicity i Human-Agent interaction. *HAI 2016 - Proceedings of the 4th International Conference on Human Agent Interaction*, 179–186. https://doi.org/10.1145/2974804.2974813

Obaid, M., Sandoval, E. B., Zlotowski, J., Moltchanova, E., Basedow, C. A., & Bartneck, C. (2016). Stop! That is close enough. How body postures influence human-robot proximity. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 354–361. https://doi.org/10.1109/ROMAN.2016.7745155

Oistad, B. C., Sembroski, C. E., Gates, K. A., Krupp, M. M., Fraune, M. R., & Šabanović, S. (2016). *Colleague or Tool? Interactivity Increases Positive Perceptions of and Willingness to Interact with a Robotic Co-worker*. https://doi.org/10.1007/978-3-319-47437-3_76

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. https://doi.org/10.1080/00140139.2012.691554

Panella, M., Marchisio, S., & Di Stanislao, F. (2003). Reducing clinical variations with clinical pathways: do pathways work? *International Journal for Quality in Health Care*, *15*(6), 509–521. https://doi.org/10.1093/intqhc/mzg057

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, *7*(3), 101–134. https://doi.org/10.1080/10864415.2003.11044275

Pfeffer, J. (2018). *The role of the general manager in the new economy: Can we save people from technology dysfunctions?* (No. 3714).

Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology*, *13*(1), 53–64. https://doi.org/10.1007/s10676-010-9253-3

Powers, E. (2017). My News Feed is Filtered? *Digital Journalism*, *5*(10), 1315–1335. https://doi.org/10.1080/21670811.2017.1286943

Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems*, *25*(4), 145–182. https://doi.org/10.2753/MIS0742-1222250405

Ragni, M., Rudenko, A., Kuhnert, B., & Arras, K. O. (2016). Errare humanum est: Erroneous robots in human-robot interaction. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, 501–506. https://doi.org/10.1109/ROMAN.2016.7745164

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Raj, M., & Seamans, R. (2019). Primer on artificial intelligence and robotics. *Journal of Organization Design*, *8*(1), 11. https://doi.org/10.1186/s41469-019-0050-0

Ramchurn, S. D., Wu, F., Jiang, W., Fischer, J. E., Reece, S., Roberts, S., … Jennings, N. R. (2016). Human–agent collaboration for disaster response. *Autonomous Agents and Multi-Agent Systems*, *30*(1), 82–111. https://doi.org/10.1007/s10458-015-9286-4

Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. *Seventh International Conference on Social Robotics*, 574–583. https://doi.org/10.1007/978-3-319-25554-5_57

Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems*, *47*(4). https://doi.org/10.1109/THMS.2017.2648849

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 101–108. https://doi.org/10.1109/HRI.2016.7451740

Rossi, A., Holthaus, P., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018). Getting to know pepper: Effects of people's awareness of a robot's capabilities on their trust in the robot. *HAI 2018 - Proceedings of the 6th International Conference on Human-Agent Interaction*, 246–252. https://doi.org/10.1145/3284432.3284464

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

Russell, S. J., & Norvig, P. (1995). *Artificial Intelligence A Modern Approach*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.8854&rep=rep1&type=pdf

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot? *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*, 141–148. https://doi.org/10.1145/2696454.2696497

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210 URL: http:/

Sanders, T. L., Wixon, T., Schafer, K. E., Chen, J. Y. C., & Hancock, P. A. (2014). The influence of modality and transparency on trust in human-robot interaction. *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 156–159. https://doi.org/10.1109/CogSIMA.2014.6816556

Sandoval, E. B., Brandstetter, J., & Bartneck, C. (2016). Can a robot bribe a human? The measurement of the negative side of reciprocity in human robot interaction. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 117–124. https://doi.org/10.1109/HRI.2016.7451742

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present and future. *Academy of Management Review*, *32*(2), 344–354.

Schwab, K. (2017). *The fourth industrial revolution*. Retrieved from https://books.google.com/books?hl=en&lr=&id=GVekDQAAQBAJ&oi=fnd&pg=PR7&dq=The+fourth+industrial+revolution&ots=NhKeFDzwhG&sig=SxKMGj8OWFndH_0YSdJMKbknCwA#v=onepage&q=The fourth industrial revolution&f=false

Shim, J., & Arkin, R. C. (2014). Other-Oriented Robot Deception: A Computational Approach for Deceptive Action Generation to Benefit the Mark. *Mobile Robot Laboratory Publications*. Retrieved from https://smartech.gatech.edu/handle/1853/52666?show=full

Shinozawa, K., Naya, F., Yamato, J., & Kogure, K. (2005). Differences in effect of robot and screen agent recommendations on human decision-making. *International Journal of Human-*

Computer Studies, *62*(2), 267–279. https://doi.org/10.1016/J.IJHCS.2004.11.003

Sierhuis, M., Bradshaw, J. M., Acquisti, A., Van Hoof, R., Jeffers, R., & Uszok, A. (2003). Human - agent teamwork and adjustable autonomy in practice. *Proceedings of the Seventh International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)*.

Strohkorb, S., Fukuto, E., Warren, N., Taylor, C., Berry, B., & Scassellati, B. (2016). Improving human-human collaboration between children with a social robot. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, 551–556. https://doi.org/10.1109/ROMAN.2016.7745172

Strohkorb Sebo, S., Traeger, M., Jung, M., & Scassellati, B. (2018). The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams. *ACM/IEEE International Conference on Human-Robot Interaction*, 178–186. https://doi.org/10.1145/3171221.3171275

Stubbs, K., Wettergreen, D., & Hinds, P. J. (2007). Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems*, *22*(2), 42–50. https://doi.org/10.1109/MIS.2007.21

Tussyadiah, I. P., Zach, F. J., & Wang, J. (2019). Do Travelers Trust Intelligent Service Robots? *Annals of Tourism Research*.

Ullman, D., & Malle, B. F. (2017). Human-Robot Trust: Just a Button Press Away. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, (March 6-9), 309–310. https://doi.org/10.1145/3029798.3038423

Ullman, D., & Malle, B. F. (2018). What Does it Mean to Trust a Robot? Steps Toward a Multidimensional Measure of Trust. *HRI 18 Companion*, 263–264. https://doi.org/10.1145/3173386.3176991

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2015). Trusting a virtual driver that looks, acts, and thinks like you. *Human Factors*, *57*(5), 895–909. https://doi.org/10.1177/0018720815580749

Von Der Pütten, A. M., Krämer, N. C., Gratch, J., & Kang, S. H. (2010). "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior*, *26*(6), 1641–1650. https://doi.org/10.1016/j.chb.2010.06.012

Walker, F., Verwey, W., & Martens, M. (2018). Gaze Behaviour as a Measure of Trust in Automated Vehicles. *Proceedings of the 6th Humanist Conference*. Retrieved from http://www.humanist-vce.eu/fileadmin/contributeurs/humanist/TheHague2018/29-walker.pdf

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, *23*(4), 217–246. https://doi.org/10.2753/MIS0742-1222230410

Wang, W., Qiu, L., Kim, D., & Benbasat, I. (2016). Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. *Decision Support Systems*, *86*, 48–60. https://doi.org/10.1016/j.dss.2016.03.007

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave new world: Service robots in the frontline. *Journal of Service Management*, (October). https://doi.org/10.1108/JOSM-04-2018-0119

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly*, *31*(1), 137–209. Retrieved from https://www.jstor.org/stable/pdf/25148784.pdf?refreqid=excelsior%3A83a8487d9380ce4c4554852ed96900d8

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *International Journal of Social Robotics*, *4*(3), 235–248. https://doi.org/10.1007/s12369-012-0144-0

You, S., & Robert, L. (2019). Subgroup Formation in Human-Robot Teams. *ICIS 2019 Proceedings*. Retrieved from https://aisel.aisnet.org/icis2019/general_topics/general_topics/18

Zhang, T., Kaber, D. B., Zhu, B., Swangnetr, M., Mosaly, P., & Hodge, L. (2010). Service robot feature design effects on user perceptions and emotional responses. *Intelligent Service Robotics*, *3*(2), 73–88. https://doi.org/10.1007/s11370-010-0060-9

Złotowski, J., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2016). Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, Journal of Behavioral Robotics, 7*(1). https://doi.org/10.1515/pjbr-2016-0005

Table 1. Main effects of dimensions on cognitive trust in AI, organized by representation

Dimensions	Robotic AI	Virtual AI	Embedded AI
Tangibility	Physical presence increases trust: More trust in robotic AI than in virtual AI.	Visual presence increases trust: More trust in virtual AI than in embedded AI.	The effect of awareness on the use of AI is not clear.
Selected references	(Bainbridge et al., 2011; K. M. Lee, Peng, Jin, & Yan, 2006; Salem et al., 2015; Shinozawa et al., 2005)	(Chattaraman et al., 2014; Mumm & Mutlu, 2011)	(Eslami et al., 2015)
Transparency	Transparency might increase trust, but the empirical research is scant.	Transparency of AI reliability and explanations on how algorithm works increase trust.	Transparency on how algorithm works increases trust; especially needed for highly intelligent managerial systems.
Selected references	(Sanders et al., 2014)	(Fan et al., 2008; Wang & Benbasat, 2007; Wang et al., 2016)	(Alan et al., 2014; Chao et al., 2016; Dietvorst et al., 2016; Dzindolet et al., 2003; Kizilcec, 2016; M. K. Lee, Kusbit, Metsky, & Dabbish, 2015; Möhlmann & Zalmanson, 2017)
Reliability	Low reliability decreases trust, but not always: When robot is perceived as having high machine intelligence - people tend to follow even a faulty robot.	Low reliability mostly decreases trust in lab and field studies where the initial trust was very high.	Low reliability significantly decreases trust and the way to restore trust is difficult and takes time.
Selected references	(Bainbridge et al., 2011; Desai et al., 2012; Freedy et al., 2007; Robinette et al., 2016; Salem et al., 2015)	(Fan et al., 2008; Glass, McGuinness, & Wolverton, 2008; Moran et al., 2013)	(Dietvorst et al., 2015; Dzindolet et al., 2003; Manzey et al., 2012)
Task characteristics	In technical tasks the trust is higher than in tasks that requires social intelligence.	In technical tasks that require data analysis trust in AI is higher than in humans.	In tasks that require social intelligence the trust in humans is higher than in AI, high self-

			confidence moderates the trust in AI.
Selected references	(Gaudiello et al., 2016; Gombolay et al., 2015)	(Ramchurn et al., 2016)	(Dietvorst et al., 2016; Logg, Minson, & Moore, 2018)
Immediacy behaviors	Responsiveness, adaptiveness and pro-social behaviors increase trust.	Personalization and use of persuasion tactics increase trust.	Personalization improves trust; constant tracking of workers' behaviors may decrease trust.
Selected references	(Baraglia et al., 2016; E. De Visser & Parasuraman, 2011; Hoffman & Breazeal, 2007; Oistad et al., 2016)	(Andrews, 2012; Fenster, Zuckerman, & Kraus, 2012; Komiak & Benbasat, 2006)	(Dzindolet et al., 2003; M. K. Lee et al., 2015; Matz, Kosinski, Nave, & Stillwell, 2017; Möhlmann & Zalmanson, 2017)

Table 2. Main effects of dimensions on emotional trust in AI, organized by representation

Dimensions	Robotic AI	Virtual AI	Embedded AI
Tangibility	Physical presence may increase liking, but also induce fear.	Presence of a "persona" increases liking and emotional trust.	Being unaware of AI use may evoke anger. Positive emotions could be driven by good reputation of a developing firm.
Selected references	(Obaid et al., 2016; Shim & Arkin, 2014)	(Chattaraman et al., 2014; de Visser et al., 2017; Pak, Fink, Price, Bass, & Sturre, 2012; Qiu & Benbasat, 2009)	(Eslami et al., 2015; Hengstler et al., 2016)
Anthropomorphism	Human-likeness mostly increases positive emotions but can also cause discomfort.	Mostly increases trust, but also creates high expectations regarding AI's abilities. Attractiveness and personalization, such as ethnicity or facial similarity to the user increases trust.	
Selected references	(Appel, Weber, Krause, & Mara, 2016; Jacq et al., 2016; Zhang et al., 2010; Złotowski et al., 2016)	(Khan & Sutcliffe, 2014; Obaid et al., 2016; Verberne et al., 2015; Von Der Pütten, Krämer, Gratch, & Kang, 2010)	
Immediacy behaviors	Human-like behaviors induce high emotional trust; Erroneous robots are liked more than flawless.	Human-like behaviors increase trust and liking, yet the effect depends on users' predispositions.	
Selected references	(T. W. Bickmore, Vardoulakis, & Schulman, 2013; Birnbaum et al., 2016; Jung et al., 2013; Mirnig et al., 2017; Sandoval, Brandstetter, & Bartneck, 2016)	(Ben Mimoun, Poncin, & Garnier, 2017; Dabholkar & Sheng, 2012; Kaptein et al., 2011; Matsui & Yamada, 2019)	

Figure 1. Trajectories of trust for robotic, virtual and embedded AI as reflected by the majority of reviewed studies