Computers in Human Behavior Reports 4 (2021) 100146

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Intelligent autonomous agents and trust in virtual reality☆

Ningyuan Sun, Jean Botev *

Department of Computer Science, University of Luxembourg, Avenue de la Fonte 6, 4346 Esch-sur-Alzette, Luxembourg

ARTICLE INFO

Keywords:
Trust modeling
Virtual reality
Agent interaction
Collaboration

ABSTRACT

Intelligent autonomous agents (IAA) are proliferating and rapidly evolving due to the exponential growth in computational power and recent advances, for instance, in artificial intelligence research. Ranging from chatbots, over personal virtual assistants and medical decision-aiding systems, to self-driving or self-piloting systems, whether unbeknownst to the users or not, IAA are increasingly integrated into many aspects of daily life. Despite this technological development, many people remain skeptical of such agents. Conversely, others might have excessive confidence in them. Therefore, establishing an appropriate level of trust is crucial to the successful deployment of IAA in everyday contexts. Virtual Reality (VR) is another domain where IAA play a significant role, yet its experiential and immersive character particularly allows for new ways of interaction and tackling trust-related issues. In this article, we provide an overview of the numerous factors involved in establishing trust between users and IAA, spanning scientific disciplines as diverse as psychology, philosophy, sociology, computer science, and economics. Focusing on VR, we discuss the different types and definitions of trust and identify foundational factors classified into three interrelated dimensions: Human-Technology, Human-System, and Interpersonal. Based on this taxonomy, we identify open issues and a research agenda towards facilitating the study of trustful interaction and collaboration between users and IAA in VR settings.

1. Introduction

The rapid growth in computational power and recent advances in artificial intelligence (AI) research and other disciplines have led to a large number and widespread use of intelligent autonomous agents (IAA). Such agents are revealed in a multitude of different forms: virtual personal assistants, for example, are already commonplace in smart devices such as mobile phones and self-driving cars (Lugano, 2017; Tulshan & Dhage, 2018); medical decision-aiding systems are, for instance, applied in nutrition (Sajeev et al., 2017) and cancer diagnosis (McKinney et al., 2020); artists utilize such agents in their creative process, e.g., for painting (Park et al., 2019) and composition (Moruzzi, 2017). Further examples for IAA can be found in various other domains, including cybersecurity (Chan et al., 2019), e-commerce (Song et al., 2019), e-sports (Vinyals et al., 2019), and manufacturing (Ghahramani et al., 2020).

Generally, contemporary IAA are designed with the intention to facilitate and improve specific aspects of their users' lives. They perceive, reason, act, react, and preferably adapt to support, interact and collaborate with humans and frequently are besting them in their designated purpose. For instance, of the above examples, self-driving cars combined with computer-controlled traffic networks have significantly decreased accident rates (Legacy et al., 2019; Teoh & Kidd, 2017). Also, medical decision-aiding agents, e.g., in breast cancer diagnosis, have already outperformed doctors (McKinney et al., 2020). In some instances, IAA prevail for reasons other than capability, such as their neutrality, due to which patients feel less judged during clinical interviews with virtual agents and, consequently, are more willing to confide in virtual agents than actual humans (Lucas et al., 2014).

However, at the same time, such agents are faced with resistance by many. Self-driving cars, for example, remain the subject of heated debate, and patients usually prefer to see a real doctor over an artificial agent (Longoni et al., 2019).

This ambivalence also manifests itself in the frequently reported concern, skepticism, and even fear toward IAA (Falco et al., 2021; Irving & Askell, 2019). For instance, despite their application being arguably still limited, people worry about being replaced by IAA in their workplaces (Bruun & Duka, 2018). An often delineated worst-case scenario is that AI can eventually reach superiority with catastrophic consequences to humankind (Brundage, 2015). Many leading AI researchers and

☆ Funding: This study has been supported by the Luxembourg National Research Fund (FNR) under grant number 12635165.
* Corresponding author.
E-mail address: jean.botev@uni.lu (J. Botev).

https://doi.org/10.1016/j.chbr.2021.100146
Received 22 June 2021; Received in revised form 15 October 2021; Accepted 18 October 2021
Available online 25 October 2021

industry experts have signed an open letter [1] to warn of this scenario and appeal to set a research agenda towards robust and beneficial AI. This has boosted research in explainable/white-box AI (XAI) (Abdul et al., 2018; Adadi & Berrada, 2018), which aims to increase the transparency and accountability of AI-based systems to promote trust in such systems.

In addition to the AI community, other scientific disciplines have also been working to facilitate more trustworthy user-IAA interaction. Virtual reality (VR) technology in conjunction with IAA is promising but still largely unexplored. Therefore, this article proposes a model for users' trust in IAA in VR based on an overview of trust-related research, which identifies and categorizes essential factors for its establishment from different scientific perspectives. After providing an overview of IAA and trust in Section 2, we discuss users' trust in IAA in VR in Section 3 and present a dedicated trust model comprising different trust dimensions in Section 4. We then discuss open issues and formulate a research agenda towards a deeper understanding of trust dynamics between users and IAA in VR in Section 5, concluding in Section 6 with a summary and outlook for further research avenues.

2. Background

Trust is an essential commodity of, for example, human-machine interaction, economic behavior, and any individual, institutional, or organizational relationship. As such, trust has been widely studied, yet mostly in specific contexts outside of VR, which, with its experiential and immersive character, allows for new ways of interaction and tackling trust-related issues. This section will first introduce the definitions and different perspectives on IAA, followed by a discussion on general trust research and, in particular, trust in virtual environments.

2.1. Intelligent autonomous agents

To discuss the definition of IAA, we must first clarify what an agent is, which is a fundamental, controversial, yet often neglected question. The debate over the definition of the term "agent" mostly appears in the 1990s when software agents began to emerge. On the one hand, several researchers tried to formalize the definition of an agent, for instance, as an object that fulfills certain users' goals (Luck et al., 1995). On the other hand, opposing voices were pointing out that the usage of the term "agent" is highly diverse to the extent that it becomes almost meaningless without specifying the context (Shoham, 1993). The current consensus on the definition in the context of agent technology is elaborated on generally in (Franklin & Graesser, 1996).

One way to define IAA is as computer agents that can perceive the surrounding environment and behave autonomously by utilizing inferential or complex computational methodologies to perform tasks for the user (Imam & Kodratoff, 1997). Specifically, an agent is considered *intelligent* if it is proactive, reactive, and social; this refers to the ability to take the initiative in order to accomplish goals, perceive and respond to the environmental changes in a timely fashion, and communicate with other entities (including both agents and humans) (Wooldridge, 1999). Also, according to (Franklin & Graesser, 1996), an agent is *autonomous* if it is "situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future." As such, a simple system like a thermostat is not intelligent but autonomous. Other, less common definitions focus on, for example, the levels of control on the behaviors of an autonomous agent, which, based on descending order of controllability, include a *motor skill level*, a *behavioral level*, a *motivational level*, and an *environmental level* (Blumberg and Galyean, 1997).

IAA are commonly investigated from either an intrinsic perspective or an extrinsic perspective. The intrinsic perspective decomposes IAA into their constituents, essentially taking a microscopic view. A considerable amount of the AI-related literature discusses IAA from this perspective, often in the context of, e.g., convolutional neural networks (Li et al., 2016) and generative adversarial networks (Goodfellow et al., 2014). The extrinsic perspective takes a macroscopic view of IAA, exemplified by the intentional systems theory (IST) (Dennett, 2009) that proposes to observe a system from a physical, a design, and an intentional stance. The physical stance is based on simple physical laws, and the object's behavior is highly predictable, e.g., a heavy object falling when letting go. The design stance is more elaborate than the physical stance and focuses on an object's particular purpose, e.g., an alarm clock setting off an alarm at a given time. While remaining predictable, actions are not as immediately consequential as in the physical stance. Finally, the intentional stance involves beliefs, desires, and rationality to accomplish specific goals, and systems such as IAA are eligible to be observed from this stance.

Both perspectives constitute an essential aspect in the design of IAA. However, for two reasons, we focus more on extrinsic aspects here. Firstly, IAA are closely integrated and able to achieve higher-order functions as an entity. People trust this entity rather than one of the specific components it is composed of. For example, a self-driving agent could incorporate components such as a computer vision system, a navigator, and a motion control system. It is trusted because of its capability to drive a car, rather than the high performance of the computer vision system or another subsystem. Another argument in favor of emphasizing the extrinsic perspective is that emotional and moral aspects play an essential role in the trust relationship between users and IAA (cf. Section 2.2.4). Likewise, XAI acknowledges trust as a psychological and sociological phenomenon and aims to bridge the gap between intrinsic and extrinsic views to promote trust in AI-based systems (Adadi & Berrada, 2018).

2.1.1. IAA interaction in VR

The interaction with IAA so far is mainly mediated via standard two-dimensional displays as in, for instance, desktops and smartphones. They are often designed in congruence with the WIMP paradigm (windows, icons, menus, and pointers) and can be interacted with via input devices such as a mouse, keyboard, or touch-sensitive screens. Immersive technologies such as virtual reality (VR) change these modes by integrating different components and sensoric systems to allow for more natural interaction, e.g., with motion and gestures. This allows users to interact with IAA similar to how they interact with other humans in the real world.

Novel VR devices can provide a high level of immersion that standard two-dimensional displays can hardly convey. However, the impact of high-level immersion on user-IAA interaction remains largely unexplored since there is only limited comparative research on the effect of different levels of immersion on user-IAA interaction. Nevertheless, studies in other contexts have shown the potential benefits and drawbacks of VR-based interaction compared to desktop-based modes. Benefits include, e.g., being more intuitive, natural, and satisfied in a game scenario (Santos et al., 2009) or enabling users to process information easier when using virtual maps (Dong et al., 2020). On the other hand, there are potential drawbacks, such that VR-based interaction may also lead to lower performance in learning tasks (Parong & Mayer, 2018) or that it may negatively impact the utilitarian value of a virtual shopping system (Peukert et al., 2019).

Haptic techniques, for instance, can be used for simulating the sense of touch and currently are the subject of extensive research in the field of VR. This line of research is usually referred to as *mediated social touch*, which affects how users interact with another entity –either a human (Cascio et al., 2019) or a virtual agent (Huisman, 2017)– in various aspects. Existing approaches to create an illusion of touch include, e.g., mid-air haptics, vibrotactile, exoskeletal, and electromechanical solutions (Price et al., 2021). These techniques are usually embedded in gloves or suits and used conjointly with immersive display technology like head-mounted displays (HMDs).

[1] https://futureoflife.org/ai-open-letter.

Another example of VR-based natural interaction effects is the *virtual hand metaphor*, which refers to a set of techniques that allow users to manipulate virtual objects "in a way that closely resembles real-world touching and grasping" (Pietroszek & Lee, 2019). While the metaphor can promote the users' perceived immersion and virtual body ownership in VR, it may lead to disadvantages such as fatigue or limited reach (Bowman et al., 2001).

2.2. Trust

The essence of trust has been subject of much scientific research. Some observe trust from a more sociopsychological perspective that emphasizes its function as a social relation. Other researchers, for example, incline towards a psychological point of view, stressing the affective basis of trust and investigating how emotions can alter trust (Dunn & Schweitzer, 2005; Myers & Tingley, 2016). There also exist works that model and quantify trust, typically in economics and computer science (Marsh, 1994). Consequently, the definition of trust diverges across disciplines. Computer scientists, economists, philosophers, psychologists, and sociologists all take their individual perspectives on trust. The need for a common definition to progress with trust research on the whole, for instance, is highlighted in (McKnight & Chervany, 1996), where the authors come to this conclusion based on a wide range of literature covering a plethora of different notions of trust.

For instance, Zaltman and Moorman (Zaltman & Moorman, 1988) define trust from a transactional, economics perspective as an "interpersonal or interorganizational state that reflects the extent to which the parties can predict one another's behavior." This implies that trust is essentially the trustors' confidence of their prediction on trustees. On the other hand, Deutsch (Deutsch, 1973) describes trust from a psychosociological perspective as "confidence that one will find what is desired from another, rather than what is feared." This definition implies that trust is driven by expectation, i.e., trust forms when a user's expectations are fulfilled. Scanzoni (Scanzoni, 1979) takes a more reward-centric perspective, viewing trust as an "Actor's willingness to arrange and repose his or her activities on Other because of confidence that Other will provide expected gratifications." This creates an abstraction of trust behaviors, which indicates the gratitude mechanism being the dominant rule of trust.

Mayer et al. (Mayer et al., 1995) proposed the current, commonly accepted definition of trust, centering around vulnerability:

> "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."

This vulnerability-based notion is a common abstract definition of trust and widely adopted to accommodate specific contexts in the literature afterward, as the following three definitions exemplify.

> "Trust is a psychological state comprising the intention to **accept vulnerability** based upon positive expectations of the intentions or behavior of another." (Rousseau et al., 1998)

> "Trust, then, on this first approximation, is **accepted vulnerability** to another's possible but not expected ill will (or lack of good will) toward one." (Friedman et al., 2000)

> "The element critically lacking in a standard PD game, as a means of studying trust, is what constitutes the core of an act of trust or trusting behavior: taking risks and thus **making oneself vulnerable**." (Cook et al., 2005)

Taking this vulnerability-based definition, together with the fact that IAA are gradually capable of accomplishing tasks that used to require highly skilled human workers, we speculate that the high level of perceived vulnerability may have prevented users from trusting IAA. The sensation of vulnerability can be further magnified if the IAA's actions touch upon critical user interests, for instance, related to health or finances.

During a user-IAA interaction, the sensation of vulnerability can originate from different aspects, which are further discussed in the remainder of this section based on various studies. Some of these studies are based on Wizard-of-Oz type experiments, which remain controversial regarding their effectiveness and research validity (Riek, 2012). However, with this article focusing on the design of trustworthy IAA, such experiments remain relevant since they demonstrate the effect of different features in IAA.

2.2.1. VR technology and trust

Until recently, the notion of "trust in information technology" has been controversial, and some researchers argued that trust only exists when the trustee has volition and moral agency, and "people trust people, not technology" (Friedman et al., 2000). This implies that trust in specific information technology, such as VR devices, is not meaningful due to the high-level compliance with users. More recent studies, however, take the opposite stance on this issue. These studies suggest that human users can be vulnerable to information technology similar to a person (Lippert, 2002; Mcknight et al., 2011), and therefore VR devices classify as a valid trustee. For example, a promising application of VR devices supports telesurgery, where doctors need to rely on the technology to operate (Javaid & Haleem, 2020).

Related literature has identified several factors that can influence users' trust in information technology. Some are exemplified in (Lippert, 2002), including technology predictability, technology reliability, and technology utility. Technology utility, in turn, includes perceived usefulness, perceived ease of use, dependence upon technology, and predilection to trust technology. Another widely accepted set of factors comprises functionality, helpfulness, and reliability, which were proposed to describe users' trusting beliefs toward information technology (Mcknight et al., 2011).

Functioning as information mediators between users and IAA, VR devices generally are sufficient for their designated purposes and only produce a minimal amount of uncertainty. Therefore, most of the time, users' fundamental reliance on the technology tends to outweigh the feeling of trust, as examined in another active line of research on the acceptance and usability of VR devices (Sagnier et al., 2020). Moreover, novel VR devices (e.g., HMDs) are typically equipped with various sensors that capture detailed information about users. This widens the scope of possible approaches that IAA or their designers can employ to establish trust with users. For instance, motion-tracking components can capture the users' body language that conveys no less significance and amount of information than verbal communication (Guerrero & Floyd, 2006). Eye-tracking components may enable IAA to infer the users' intentions by applying the theory of mind on eye-gaze data (Narang et al., 2019). Such extended capabilities can also generally alter the trust formation process between users and IAA in VR.

Nevertheless, VR devices are potentially riskier to their users than other common information mediators such as smartphones (Adams et al., 2018). For instance, data captured by the sensors embedded in VR devices can be sensitive (e.g., eye and body motion trajectories), leading to more invasive personal inferences once compromised. Users can also be exposed to harassment and other crimes when immersed in virtual environments, especially when embodied in human-form avatars.

Notably, a few recent studies (Salanitri, 2018; Salanitri et al., 2015, 2016, 2020) have focused on trust in VR technology and explored its association with two major aspects of the interaction experienced via immersive VR technologies: the level of presence and system usability. Both aspects were found to significantly correlate with trust in VR technology and, in turn, can be impacted by the parametric features of VR devices, such as refresh rate and field of view (more features are discussed in (Schuemie et al., 2001)).

2.2.2. Automated systems and trust

By interacting with automated systems like IAA, users typically are already putting some degree of trust in them since uncertainty in such systems always exists, especially when considering future events, where absolute certainty can only be approximated. However, such degree of trust is not necessarily appropriate, which may lead to inconsistent trusting behavior, i.e., users sometimes putting too much, and on other occasions not enough, trust in automated systems like IAA. This is a known phenomenon that can be captured as the relationship between trust in and the capabilities of automated systems (Lee & See, 2004) as illustrated in Fig. 1. The diagonal indicates so-called *calibrated trust*, i.e., an appropriate, well-balanced trust level set between trust and trustworthiness. *Overtrust* denotes a situation where trustors' trust exceeds the trustees' capabilities. *Distrust*, in turn, is when trustors underestimate trustees. Overtrust and distrust lead to misuse and disuse, respectively, and consequently can decrease user-agent interaction efficiency.

The *resolution* shown in Fig. 1 captures the potential variances in the range of trust. The central resolution area shows a range of trustworthiness mapping onto the same range of trust, denoting the optimal situation. In contrast, the inner resolution area indicates a larger range of trust mapping onto a lesser range of trustworthiness and, respectively, the outer resolution area indicates a larger range of trustworthiness mapping onto a lesser range of trust.

This calibration-based model, however, represents trust mainly from an information-processing perspective. Automated systems in this model play the role of an information source. Human users observe the source, i.e., automated systems, from which relevant information, such as their capabilities or reliability, can be obtained as a basis for users to form trust. Automated systems in this model always function in their own way without considering how trustworthy they may appear. Consequently, users have to adjust their expectations in automated systems to form calibrated trust, which is common in operating supervisory control automation.

A more recent trust model argues that automated systems driven by novel AI algorithms can actively adjust their perceived trustworthiness to help calibrate trust (Chiou & Lee, 2021). According to this model, trust goes beyond unilateral information processing and becomes relational as both parties –for instance, human users and IAA– can actively adapt to the other side.

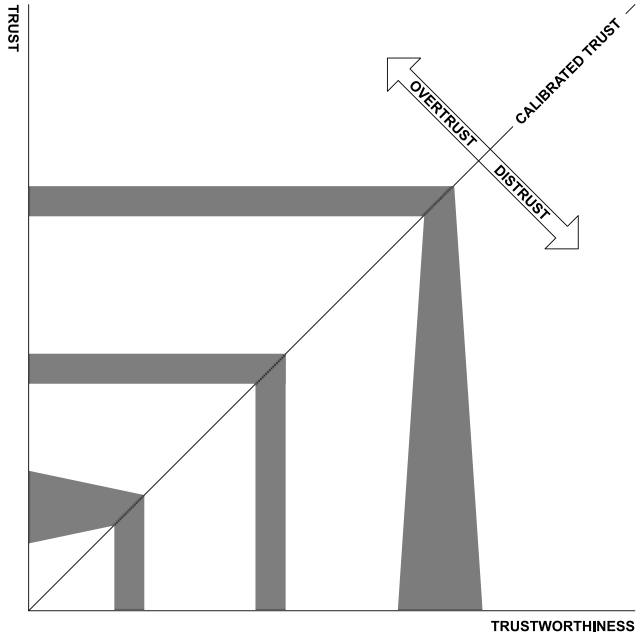

Fig. 1. Trust behaviors (adapted from (Lee & See, 2004)). The horizontal and vertical axes represent the trustworthiness of IAA and users' trust in IAA, respectively. The diagonal stands for the case where users put an appropriate level of trust in IAA. The areas over and below the diagonal depict the overtrust and distrust situation, respectively. The gray area is the resolution.

A widely accepted theory of trust in automated systems (Muir, 1994) is extended from (Rempel et al., 1985), which models interpersonal trust as three dimensions: *predictability*, *dependability*, and *faith*. When applied to automated systems, predictability refers to how predictable the recurrent behaviors of automated systems are. It is regarded as the basis of and positively correlated with trust in automated systems. Three more specific factors comprise the main determinants of predictability, namely the actual predictability, the users' ability to estimate the predictability, and environmental stability.

The actual predictability is inversely associated with the constraints on the behaviors of automated systems. IAA as autonomous systems are programmed to be less restricted on their behaviors, regardless of whether in VR or not. Therefore, the actual predictability of IAA's behaviors is inherently low, which is a barrier to establishing trust in IAA. Even for those fairly predictable IAA, their predictability can still be underestimated due to the users' impression of lacking constraints and supervision on these IAA. Moreover, contemporary IAA with deep learning techniques are often too complex to be fully understood, which also prevents users from inferring the behaviors of IAA. As such, users can only rely on empirical observations to estimate IAA's predictability, which is impractical in critical situations where the behaviors of IAA may lead to severe consequences.

Being aware of these issues, IAA designers also endeavor to facilitate users' estimations by improving the transparency of IAA. VR is particularly advantageous to this process because many limitations (e.g., physics, costs, and technology) constraining IAA in reality can be removed in virtual environments. One effective application of this advantage is allowing users to conduct simulations with IAA in virtual environments before deploying them.

The stability of the environment that an automated system interacts with can also improve the predictability of the system. This is exemplified by IAA's outstanding performance in playing different kinds of video games, including real-time strategy games like StarCraft (Vinyals et al., 2019), first-person shooter games such as Doom (Lample & Chaplot, 2017), and many more. Compared to the real world, these virtual environments are usually less unstable and noisy and, therefore, can significantly facilitate IAA's performance. Conversely, IAA can hardly maintain the same level of competence in real-world scenarios, such as autonomous driving.

Predictability is primarily decisive in trust early during the interaction. Afterward, the impact of predictability fades, and dependability becomes more relevant instead. The dependability of an automated system refers to the extent to which the system can be relied upon by its users. Throughout the interaction, users gather evidence of system performance, transforming uncertainty (predictability estimation) into certainty (empirical observations). Based on the obtained experience and knowledge, users attribute dependability to automated systems. Acquiring such experience and knowledge may demand trials, which, as previously mentioned, can be resource-consuming, destructive, or unattainable in reality, whereas in VR, it can be drawn from simulations with IAA.

Faith denotes users' confidence in the consistency of systems' future behaviors. By having faith in an automated system, users are confident that the system will be as predictable and dependable as it currently is. Faith not only matters in trust in automated systems or trust in IAA but also underpins trust in all other contexts since the future is always unpredictable, regardless of how close it is to the present moment or how predictable we think the future is.

Another acknowledged theory has also identified three key bases of trust in automated systems: *performance*, *process*, and *purpose* (Lee & Moray, 1992). Performance is users' expectations of consistent, stable,

and desirable behaviors of automated systems. Process refers to the algorithms or rules that automated systems are compliant with. Purpose describes the designers' intentions reflected in automated systems. However, for two reasons, this theory is not fully compatible with user-IAA interaction. Firstly, contemporary IAA usually employ so-called black-box algorithms that even specialists in relevant fields may find challenging to comprehend, leading to the reduced influence of process. Secondly, IAA are discarding preset rules that reflect programmers' intentions and given more freedom and authority on decision making, which decreases the influence of purpose.

Trust studies that take a context-independent stance are relatively rare compared to the substantial context-specific trust research. Nonetheless, the accumulation of the latter may outline a ubiquitous theory of trust and inspire future research on users' trust in IAA in VR. The following discussion primarily draws from three studies (Hoff & Bashir, 2015; Lee & See, 2004; Schaefer et al., 2016) that provide an overview of context-specific trust research on automated systems. From these studies, we isolated several significant factors that impact trust in automated systems in addition to what has been covered by the above-mentioned theories.

Visual design has an overarching effect on users' trust in automated systems like IAA. For users, visual information is pronounced and accounts for a large fraction of the information perceived, especially at the beginning phase of the interaction when users have only minimal knowledge to evaluate the trustworthiness of a system. Various studies revolve around the influence of visual design on trust spanning from simply altering the color and layout of basic widgets (Wakefield et al., 2004) to utterly different art styles (McDonnell et al., 2012; Yeh & Wickens, 2001). Different agent representations also impose different impacts on trust (Nowak & Biocca, 2003).

System failure refers to an automated system's failure to accomplish their tasks. Its influence on trust is generally detrimental but to different extents depending on several factors as in, for instance, task difficulties (Madhavan et al., 2006), failure types (Johnson et al., 2004), and the timing of failure occurrence (Manzey et al., 2012). Trust-repairing measures after trust violations, such as explaining to users after a system failure (Dzindolet et al., 2003), can increase trust resilience. Conversely, successful user-system interaction is also interesting but insufficiently researched. For IAA, this is likely because they are still not competent enough in most cases, and, consequently, failures appear as a more prominent issue than success.

Agency of different degrees may have different influences on trust in automated systems. A study shows that users prefer adaptive agency systems that can automatically coordinate different levels of assistance among no aid, soft aid, soft intervention, and hard intervention (Cai & Lin, 2012). People also prefer to delegate to software agents when they feel in control of the agent (Stout et al., 2014).

Explainability is a concept similar to transparency, emphasizing the system's capability of taking the initiative to make itself understandable. Recently, there has been an upsurge in explainability research in the AI community, and increasing trustworthiness is defined as one of the various purposes for developing explainable AI (Arrieta et al., 2020).

While the trust-related theories and findings mentioned so far are mostly oriented toward dyadic interaction, there also exists substantial research on the trust relationships among multiple entities, especially in the field of multi-agent systems. Analyzing an agent's trustworthiness in multi-agent situations is complex since its trustworthiness is not only determined by itself but also influenced by its connections to other agents, systems, or users. In social psychology, for instance, such relations have been modeled as sentiment networks and shown to not remain stable but tending to disintegrate (Cartwright & Harary, 1956; Leskovec et al., 2010). The separate representation of a single IAA's components as if there were several agents further increases the complexity (Baylor, 2009). Trust in such cases is fuzzy since it is perceived as a combination of several trustees.

When users interact with an unknown agent or cannot confidently evaluate its trustworthiness from the past interactions with it, third-party testimony –i.e., deriving an agent's trustworthiness from the agent's past interactions with others– can be central to judge its trustworthiness (Yu et al., 2013). Eventually, this will result in users and agents forming a trust network whose complexity requires reliable analytic methods and mathematical models to study the underlying dynamics similar to those in computational trust (Marsh, 1994).

Besides user-agent trust, inter-agent trust plays an essential role in multi-agent systems and is, for instance, the subject of various recent studies on trust management within the Internet of Things (IoT) (Din et al., 2021; Fortino et al., 2020, 2021; Hriez et al., 2021). The primary focus here lies on modeling inter-node (i.e., inter-agent) trust to maximize the performance and efficiency in an IoT system consisting of multiple edge devices. In particular, current edge devices are often considered IAA due to their ability to perform AI-driven and self-* tasks locally on their embedded computational and perceptual units. As such, users' trust in IAA can also be subject to the adopted trust management scheme since it may substantially define IAA in terms of their capability, reliability, efficiency, and similar factors.

An IoT system itself can be regarded as an intelligent adaptive agent. However, trust in entire IoT systems remains under-researched and a complex issue (Aldowah et al., 2021). Some recent studies have investigated and promoted trust in IoT systems, for example, by employing explainable AI algorithms (García-Magariño et al., 2019). From a broader perspective, trust in IoT systems can also be tackled by considering users, services, and edge devices equally as agents, which together form a larger IoT network (Kravari & Bassiliades, 2019).

The discussion around trust becomes even more complex when involving distrust. Despite the terms "trust" and "distrust" appearing as linguistic opposites, their very substance can lie in different spheres. The notion of differentiating trust from distrust has recently been increasingly accepted in the scientific community, acknowledging that while the constructs of trust and distrust indeed overlap to some degree, they also differ substantially. For instance, while the reliability and credibility of an information system have a critical impact on both trust and distrust, the experienced strain shows more relevance to distrust, whereas well-being and performance were found to be more connected with trust (Thielsch et al., 2018). Likewise, the users' trust in a website is found to be mainly influenced by social factors like recommendations from friends, while users' distrust is primarily produced from the website's structural design, such as pop-ups or a complex layout, and both can be affected by the contents of the website (Seckler et al., 2015).

2.2.3. *Virtual environments and trust*

The Internet, VR, and related technologies have been rapidly developing over the last twenty years, during which many conventionally offline activities have migrated into virtual environments as in, for example, virtual conferences, virtual chat rooms, and massive multiplayer video games. With an increasing number of interpersonal and user-agent interactions now happening virtually, the trusting behaviors inside these virtual environments are drawing attention and, consequently, have sparked various related studies in the field of agent design, human-computer interaction, sociology, psychology, and more. Many of these studies have adopted the vulnerability-based trust definition, indicating that the nature of trust within virtual environments is essentially the same as in reality. However, the anonymity of user identity and designers' complete control in virtual environments has allowed and led to many interesting findings.

Trust in virtual agents, for instance, is one of the most frequently studied topics, and many factors were tested on their influence on trustworthiness. For example, an agent mimicking users' behaviors, such as head movement, is rated more trustworthy than those agents that do not (Bailenson and Yee, 2005). Anthropomorphism is demonstrated to enhance trust resilience –i.e., the trust recovery process after trust violation– to a virtual agent (De Visser et al., 2016). Showing users the highlight moments of an agent, i.e., the agent's actions in key

situations, was found to significantly increase the likelihood of the agent being delegated a particular task (Amir & Amir, 2018). In human-agent teamwork, a shared mental model was found to be essential to establishing trust between the hybrid team members (Hanna & Richards, 2018).

Aside from virtual agents, also avatars, i.e., the embodiment of users in virtual environments, are commonly studied in conjunction with trust. For instance, as opposed to no representation, a self-avatar was reported to promote trust formation with others (Pan & Steed, 2017). In a study investigating how elderly users interact with others through avatars, age similarity was found to boost the trusting behaviors of the users toward their online partners (Lee et al., 2018). The embodiment of users in virtual environments significantly impact user-IAA interaction as exemplified by the Proteus Effect (Yee & Bailenson, 2007). People behave differently depending on how they are represented in virtual environments and tend to "conform to the behavior that they believe others would expect them to have" (Yee & Bailenson, 2007). As such, the impact of embodying users differently on trust can be potentially huge and needs further exploration, especially in immersive virtual environments such as VR. We regard this as an open issue, which is discussed in detail in Section 5.3.

In virtual environments experienced via immersive VR technologies, further aspects gain significance (Perkis et al., 2020). For example, trust in autonomous vehicles is often studied in simulated 3D virtual environments, the results of which can benefit the design and validation of real-world autonomous vehicles as a product (Morra et al., 2019). Also, social and psychological experiments revolving around trust are facilitated by the high level of immersion and controllability of virtual environments (Pan & Hamilton, 2018), as exemplified by (Hale & Antonia, 2016) where the delay of mimicry behavior toward participants can be precisely set to a certain number.

2.2.4. *Agents' human semblance and trust*

With the appearance and capabilities of virtual agents approaching those of actual humans, trust in IAA may move beyond trust in automated systems and instead increasingly resemble actual interpersonal trust. Many studies show that people interact with computers also in a decidedly human-like or pro-social fashion, especially when computers simulate and exhibit social cues, including, e.g., gender, ethnicity, or inner-group features (Nass & Moon, 2000). Such a tendency to interact with agentic objects in a social manner is regarded as one of human's innate abilities (Takayama, 2009).

The fundamental theory that underpins the impact of anthropomorphism on user-agent interaction was proposed as the computers-are-social-actors (CASA) framework, which states that people automatically and unconsciously apply social rules and norms to computers (Nass et al., 1994). Over the years, a large number of studies have emerged under the CASA framework, identifying numerous computer behaviors to which users respond socially or emotionally, such as praising (Mishra, 2006), blaming (Mishra, 2006), and flattering (Lee, 2010). Reciprocity (Fogg & Nass, 1997; Moon, 2000), similarity attraction (Moon & Nass, 1996), and social categorization (Nass et al., 1997) were found to also exist in human-computer interaction.

Some studies have demonstrated the positive influence of adopting anthropomorphic features on trust in virtual agents. For instance, virtual agents with a human-semblance representation were found to possess a higher level of trustworthiness than those agents that are represented in other forms (Philip et al., 2020; Weitz et al., 2019). Specifically, facial appearance and expressions, of which a subtle change may lead to very different social cues, can also influence the trustworthiness of a human-semblance virtual agent (de Melo et al., 2014; Krumhuber et al., 2007; Todorov et al., 2015). The attractiveness of facial appearance may be equally or even more important than reliability (Yuksel et al., 2017).

Verbal communication can convey social cues of trustworthiness, as well. For example, apologizing to users after system failure can strengthen trust resilience and dampen the negative impact of trust violations (De Visser et al., 2016). In an experiment of taking courses and exams with co-learner agents, users put more trust in those agents that exhibit caring behaviors than those that did not, such as complimenting and encouraging (Roselyn Lee et al., 2007). A virtual agent's voice per se can also influence its trustworthiness (Nass & Lee, 2001), despite being limited (Torre et al., 2019).

Novel VR devices provide a more natural and immersive user-IAA interaction, during which IAA's anthropomorphic features may appear more believable and induce stronger responses from users. Nonetheless, some studies have revealed the downside of using anthropomorphic features. For example, inconsistent appearance and voice (e.g., an agent with a human face but with a synthesized voice) can decrease an agent's trustworthiness (Gong & Nass, 2007). The so-called Uncanny Valley (Mori, 1970) issue also increases the risk of using anthropomorphic features on IAA. This Uncanny Valley effect was also found when virtual agents possess near-human-level intelligence (Stein & Ohler, 2017).

Besides, social norms differ among cultures and, thus, the same anthropomorphic feature may be interpreted differently when IAA are faced with multicultural user groups. For example, the number "666" represents evil and the satanic in the western hemisphere, whereas it is considered a lucky number in Chinese culture. Japanese usually express themselves in relatively polite and humble language using honorifics compared to western interpersonal communication (Ide, 1982). These differences in interpersonal interaction may lead to different trust formation conditions and processes.

Other than anthropomorphic features, the social context of the user-IAA interaction can also be decisive in the formation of trust. This includes, for example, peer-like relationships in a hybrid team of humans and IAA, delegatory relationships where users pass on their controllability over a task to IAA, tutoring relationships where IAA train users to operate complex machinery (Rickel & Johnson, 1999), companion relationships where autonomous virtual pets help to cure mental disruption (Nakajima & Niitsuma, 2020), advisor-advisee relationships where IAA recommend shopping items or medical decisions, and many more. The formation of trust can be very different across these contexts for various reasons, including different levels of accountability (e.g., users may risk their lives when using autonomous cars, whereas users risk almost nothing when interacting with a virtual pet). Shaping these social relations, in turn, can also be influenced by anthropomorphic features, such as IAA's way of communicating, politeness (Parasuraman & Miller, 2004), etc.

3. User-IAA trust relationship

While interpersonal trust is usually bidirectional, the user-IAA trust relationship comprises only a unidirectional trust relation of users toward IAA. The opposite direction, i.e., IAA's trust in users, is technically invalid since the sensations of vulnerability and willingness are the standing terms for human mental states and can be used on IAA only in an analogical way. IAA's "trust" in users is mostly manifested mathematically as in the studies of computational trust, affective computing (Picard, 2000), and the theory of mind (de Weerd et al., 2017). In this discussion, we take a user-centered perspective and focus on users' trust in IAA. Unless specified otherwise, all trust relationships mentioned in the rest of this article refer to users' trust in IAA in VR settings.

Trust in IAA forms in all user-IAA interactions and is mainly sourced from the extent to which users are uncertain about, for example, the actions of IAA and the outcomes of such actions. When users are entirely certain about the actions of IAA and the consequences of such actions, the notion of trust is hardly involved here as users know what will happen; conversely, when users are entirely uncertain, such a relationship can be described more accurately as hopeful. Trust can thus be regarded as the state of being in limbo between complete certainty and uncertainty. Certainty and uncertainty can co-exist, i.e., users may be certain of some aspects but uncertain about others.

From an average user's perspective, the predominant uncertainty

toward an agent can be summarized into one question: "can this agent achieve my expectations?". Despite seemingly straightforward, such expectations in fact constitute many smaller, specific expectations in various aspects of IAA (e.g., their capabilities, responsiveness, or aesthetics), some of which users may not even be consciously aware of. Consequently, the uncertainty underlying trust in IAA can be further divided, as well. Here, we approach them via the IST (cf. Section 2.1). Corresponding to the physical, design, and intentional stance, uncertainty about IAA is mainly produced on three levels of functionality: technical, automatic, and autonomous.

On the technical level, uncertainties are sourced from the VR devices by which user-IAA interaction is mediated and facilitated, for example, computers, HMDs, and other peripheral devices. In most cases, these VR devices perform stably and produce only a minimal amount of uncertainty, which users often ignore. Nonetheless, there are exceptions where the uncertainty of VR devices is crucial such as system latency in time-critical tasks.

The automatic level is built upon the functionality provided at the technical level and extends the functionality to the level of an automated system. On this level, the algorithmic aspects of IAA stand out over others, and IAA are observed as designed systems that adhere to a set of rules and react accordingly. From an extrinsic perspective, the functionality on this level can already be deemed as "making decisions", although such a decision is the product of enforcement and only reflects the programmers' intents. For instance, while a virtual personal assistant can be known for its highly autonomous functions, it also has automated functions such as alerting users of their daily events 15 min in advance, like clockwork. The uncertainties at the automatic level are usually manifested as, e.g., malfunction or inefficiency.

The autonomous level, in turn, is based on the automatic level. IAA on this level are observed as fully autonomous systems, i.e., systems that make their own decisions when there are numerous available choices that are not covered by preset rules. The most concerning uncertainty toward IAA on this level is agents' capability. The automation-related uncertainty, such as malfunction and inefficiency, can also be produced on the autonomous level since autonomy is essentially a high level of automation. Besides, autonomous agents may be perceived as intentional (i.e., having "volition" to choose from several choices), producing additional uncertainty related to social cues such as benevolence.

At each level, users form trust to cope with the uncertainty produced at that level. This results in three relatively distinctive trust relationships: *trust in technology*, *trust in automation*, and *trust in autonomy*. The three trust relationships essentially add up to the trust in IAA.

Each trust relationship can be further divided into dispositional, situational, and learned trust according to the trust analysis framework proposed in (Marsh & Dibben, 2003). Dispositional trust constitutes a personal trait and refers to trustors' overall trusting tendency. It is mainly determined by the factors directly related to trustors, such as their backgrounds, beliefs, and personalities. Dispositional trust typically remains stable in the short term, exerting a general impact on the user-IAA interaction. Situational trust rises from the context in which the interaction occurs (e.g., task difficulties, task contents) and, therefore, can be dynamic throughout the interaction in light of contextual changes. Learned trust is formed based on the experience in the past interaction with the same agent or other similar agents. Essentially, the formalization of learned trust is a process of users familiarizing themselves with IAA. The acquired certainty can be the evidence of either trustworthiness or untrustworthiness, leading to trust increase (or distrust reduction) or distrust increase (or trust reduction), respectively. Over the course of the interaction, learned trust usually becomes more decisive than dispositional trust and situational trust.

Dispositional trust and situational trust are primarily determined by the factors that originate from the users themselves and the contexts in which they are embedded. To some extent, IAA designers may predict and cope with different users and contexts, yet not all situations can be foreseen and dealt with at design time, making IAA less controllable by their designers.

We will cover dispositional trust and situational trust in Section 3.1 and Section 3.2, respectively. Due to learned trust being particularly relevant and central to most trust research, it will be discussed in more detail in Section 4.

3.1. Dispositional trust

Research on dispositional trust in a specific VR technology is rare, yet two other types of research can relate to this issue. The first type of research focuses on users' trust-based behaviors (e.g., accepting and using) toward a specific VR device. For example, a study that reveals the gender difference in the perceived usability of HMDs (Felnhofer et al., 2013) may also indicate the gender difference in the trusting attitudes to HMDs. Another type of research investigates users' trust in other technologies. For instance, people of different age groups are estimated to hold different levels of trust in Internet technology (Blank & Dutton, 2012), which may also be the case for VR or related technologies.

Users' dispositional trust in automated systems is found to be influenced by four major factors: culture, age, gender, and personality (Hoff & Bashir, 2015). For example, compared to Americans, Mexicans tend to put more trust in automated decision-aiding systems and less in manual systems (Huerta et al., 2012). Younger and older users exhibited different levels of trust in a diabetes management system that shows a picture of a physician on its user interface (Pak et al., 2012). Females have rated the same security robot as more trustworthy than males have (Gallimore et al., 2019). Neuroticism, one of the Big Five higher-order personality traits, is negatively correlated with users' acceptance of the advice given by automated systems (Szalma & Taylor, 2011).

The same social cue of trustworthiness may also have a different effect on different dispositions. For instance, when exposed to the same facial cues that convey untrustworthiness, older adults find the face more trustworthy than younger adults (Castle et al., 2012). This implies that IAA with human-semblance representation may appear trustworthy to different extents depending on their user group.

Dispositional trust is clearly tangible in single-user scenarios, whereas different users may have different dispositions in groups of users, making this a more complex issue. For instance, the variety within a user group, especially when the group was simply divided based on demographics such as age or gender, can lower the efficacy of a particular IAA design targeting the user group (Righi et al., 2017). For IAA designed for more general purposes, such as conversational agents, there can be several clusters of users showing comparatively distinct overall dispositions. Increasing the trustworthiness of such agents requires their designers not only to consider the main dispositions exhibited among all the users but also the varieties within each user group.

3.2. Situational trust

Depending on the context by which a user-IAA interaction is framed, the user's trust may appear differently. This is reflected in the diverse scenarios in which trust-related issues have been explored, including, e.g., educational, clinical, and organizational contexts. As such, trust in a specific entity or individual may also fluctuate over time due to contextual changes. For instance, voters' trust in a candidate can swing throughout the election campaign due to, e.g., geopolitical turbulence.

However, considering context is not straightforward since it is highly abstract and involves many instances. Each instance, no matter how minimally it differs from others, potentially has its own specific influence on the trust relationship that the context encompasses. Moreover, users and IAA as entities themselves can also influence situational trust since some aspects are unrelated to the interaction per se and belong to the context. For example, the positive mood of an individual is found to boost the initial trust in an automated decision-aiding system (Stokes et al., 2010). However, the positive mood is neither the product of the

individual's background nor directly related to the system; it instead may result from how the individual's day progressed so far.

Hence, minimizing the influence of situational trust can reduce the complexity of user-IAA trust research specific to VR. Such minimization can be approached either by strictly limiting the study to a specific context or incorporating as many scenarios as possible to exploit the factors that exert a similar impact across scenarios. Nevertheless, both approaches have their drawbacks. Focusing merely on a specific context limits the possibility of generalizing the study into other scenarios. Considering multiple scenarios requires considerable time and effort, which can be a massive impediment for a single piece of trust research.

There exist efforts to converge the current discrete trust studies across scenarios. According to (Hoff & Bashir, 2015), the factors that influence situational trust in an automated system can be divided into two categories: external variability and internal variability. External variability includes the system type, system complexity, task difficulty, task uncertainty, workload, distractor, organizational setting, and the framing of a task. Internal variability includes self-confidence, subject matter expertise, mood, and attentional capacity.

Among the many possible contexts, in particular, the multi-agent situation stands out. Taking the perspective that trust constitutes a one-to-one relationship, the existence of other agents can be regarded as part of the context. It has been shown that the presence of another agent alone is sufficient to alter the perceived trustworthiness of an agent and that a relatively unreliable agent is rated more trustworthy when paired with a reliable agent compared to when paired with the same unreliable agent (Ross, 2008). Third-party testimony, as mentioned in Section 2.2.2, is another example of how the formation of trust in an agent can differ in multi-agent contexts. Inter-agent trust is beyond the scope of this article, with multi-agent systems research mainly focusing on discrete computation- and encryption-based protocols (Pliatsios et al., 2020).

4. Trust dimensions

Section 2 and Section 3 discussed a large number of individual factors that can impact trust in IAA. Combining multiple factors, however, these can play a role at different times during the interaction or mutually influence each other, which considerably complicates the analysis and design of trust in user-IAA relationships in VR. To be able to assess and understand the various and often intertwined factors, we propose an abstract, layered model for learned trust between users and IAA, which comprises a *Human-Technology* dimension, a *Human-System* dimension, and an *Interpersonal* dimension. Each dimension encompasses different factors that influence trust, and the overall impact of each dimension on trust is relatively distinct from the perspective of engineered systems.

4.1. Dimension structure

The Human-Technology dimension mainly consists of factors related to the information mediators, i.e., VR devices, as outlined in Section 2.2.1. These factors are highly relevant to the uncertainty produced on the technical level and, therefore, mainly influence the trust in technology.

The Human-System dimension contains factors related to the automated or autonomous aspects of IAA (cf. Section 2.2.2). The impact of this dimension is determined by the uncertainty generated at the automatic and autonomous level and, therefore, is mainly relevant to the trust in automation and the trust in autonomy. The Human-System dimension is arguably the most crucial to the user-IAA trust relationship today, since most of the problems with IAA concern system properties, such as their incompetent capabilities, poor stability, etc. Moreover, current IAA usually are proficient at handling only one specific task but incapable of others. While there is a lack of generalizability, IAA are thus more of an automated system than a mere information system. However, they are not at the level of an independent, animated social actor, despite resembling humans in intelligence, behaviors, and appearance.

The Interpersonal dimension comprises factors related to interpersonal trust relationship, as discussed in Section 2.2.3 and Section 2.2.4. The impact of this dimension is not directly relevant to the uncertainty produced at any level but instead results from our innate ability and tendency to interact with everything around us socially, therefore affecting all three trust relationships. The Interpersonal dimension will likely become the most relevant dimension in the next few decades owing to the developing capability of IAA. Highly reliable IAA only produce a minimal amount of uncertainty at the automatic and autonomous level, resulting in the Human-System dimension being insignificant to trust. The beginning of this millennium had already witnessed a similar shift –as exemplified by the CASA framework and affective computing proposals– when scientists started to focus on the social and emotional aspects of computers.

It should be noted that this tripartite dimensional structure is only one possible categorization of the factors. For instance, taking a broader perspective, the Human-Technology and Interpersonal dimensions could be embedded into the Human-System dimension to encompass all the aspects. However, the three dimensions we propose here offer a more unambiguous indication of how to analyze and design trustworthy IAA. Nevertheless, although discussed individually, their boundaries are fuzzy, and the dimensions are interrelated, determining the users' trust in a dynamic, collective way. We will discuss these interrelations and dynamics in the following Sections 4.2 and 4.3.

4.2. Dimension interrelations

To clarify their interrelations, we must first classify the trust-related factors. All the factors identified to influence trust can generally be divided into primary and intermediate factors. *Primary factors* are essentially the objective properties of an entity (trustors, trustees, or the environments) and can be accurately measured. Examples of primary factors include specific parameters of the agent's algorithms, the pupil's color of a human-semblance virtual agent, the response time of the controllers for VR devices, and many more. *Intermediate factors* are abstract properties or subjective evaluations that are attributed to an entity. Intermediate factors are not directly measurable and can only be indicated via primary factors. Examples of intermediate factors include agent capability, agent reliability, aesthetics and agent representation, agent credibility, the stability of the environment, the level of presence, and the general trusting attitudes of trustors.

As Fig. 2 illustrates, each intermediate factor can be influenced by one or more primary factors and, in turn, can influence one or more trust dimensions. Thus, altering one primary factor may influence more than one dimension, manifested as the interrelations among the dimensions.

The interrelations can also be reflected in the theories proposed to explain different types of trust (e.g., human-technology trust, human-automation trust, human-computer trust, and interpersonal trust). Although aiming at different subjects, these theories constitute similar structures. Ability, benevolence, and integrity, which were originally employed for explaining interpersonal trust, are also adopted to address other types of trust, such as trust in technology. Reliability, dependability, and faith, which were initially introduced to describe human-system trust, in turn, can correspond to ability, benevolence, and integrity, respectively. The three dimensions might be governed by the same mechanics, which is also suggested in (Dennett, 2008), which states that people have formed an evolutionary strategy to deal with all the entities encountered, including humans, animals, and artifacts, as if they were a rational agent with beliefs and desires.

4.3. Dimension dynamics

The influence of each dimension on trust is dynamic throughout user-IAA interaction. To explain their dynamics in detail, we need first to

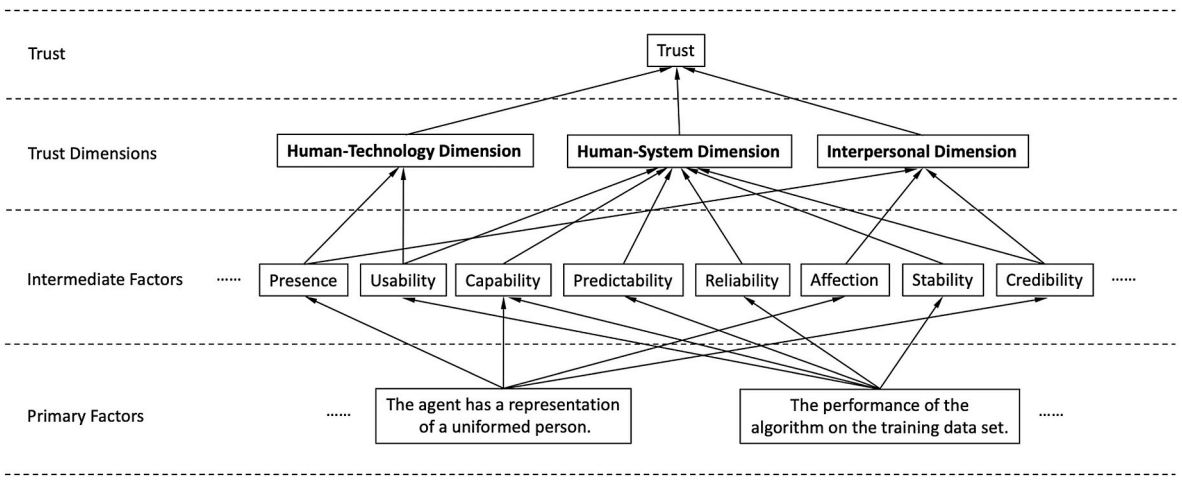

Fig. 2. The organization of trust-related factors and trust dimensions. An arrow illustrates the influence of one on the other. While the first two intermediate factors, i.e., presence and usability, are relevant to VR devices, the rest is relevant to IAA.

specify users and contexts because the dynamics can be different depending on these two parameters (cf. Section 3.1 and Section 3.2). Fig. 3 exemplifies the potential dynamics of two users' trust in a single intelligent autonomous agent in three different contexts. The three subfigures, i.e., Fig. 3a, Fig. 3b, and Fig. 3c, illustrate the influence of the trust dimensions on trust at three different times t_0, t_1, and t_2, respectively. The time t_0 marks the beginning of the user-agent interaction, while t_1 and t_2 are some later points in time with t_1 being closer to t_0. In each subfigure, the rectangular plane illustrates the situation-disposition space that is defined by two orthogonal axes. The z-axis denotes all possible dispositions, while the x-axis denotes all possible situations. As such, a point (e.g., $S1D1$) on the plane symbolizes a specific disposition or user ($D1$) in a specific situation ($S1$). Two different points on the same horizontal line (e.g., $S1D1$ and $S2D1$) refer to the same disposition ($D1$) in two different situations ($S1$ and $S2$). Likewise, two different points on the same vertical line (e.g., $S2D1$ and $S2D2$) represent two different dispositions ($D1$ and $D2$) in the same situation ($S2$). Two different points that are neither on the same horizontal nor on the same vertical line (e.g., $S1D1$ and $S2D2$) represent one disposition ($D1$) in one situation ($S1$) and the other disposition ($D2$) in the other situation ($S2$).

At each point (e.g., $S1D1$), there is a bar illustrating trust or distrust of that specific user ($D1$) in the agent under that specific situation ($S1$). While distrust is illustrated as a red, below-the-plane bar, trust is represented by an above-the-plane bar, which, in turn, comprises maximally three smaller, colorful bars (green, yellow and blue). The volumes of the three smaller bars represent the relative influence of the three dimensions on trust. The height of the trust bar and the distrust bar means the level of trust or distrust.

Here, to facilitate the later discussion, we introduce an example context of the agent playing board games for its users. Assume that: (1) The agent can play numerous board games, and its performance correlates to the complexity of the board game (e.g., proficient at Gomoku[2] but incapable of Go); (2) The agent has a cartoonish cat appearance in VR; (2) $S1$ denotes the situation where users play Go against an opponent and want to delegate this game to the agent; (3) $S2$ denotes the situation where users play Gomoku against an opponent and want to delegate this game to the agent; (4) $S3$ denotes the situation where users play a niche, child-oriented board game against an opponent and want to delegate this game to the agent; (5) $D1$ denotes a thirty-year-old user named Alpha who has never used any VR devices before; (6) $D2$ denotes a five-year-old user named Beta who has never used any VR devices either.

As such, the bars at $S1D1$ across all three subfigures illustrate the dynamics of Alpha's trust in the agent in the context of the agent playing Go on behalf of Alpha. Alpha initially (cf. Fig. 3a) has a certain level of trust in the agent, believing that it has a fair chance to win the game. The three dimensions may contribute equally to the initial trust due to Alpha's unfamiliarity with both the VR device and the agent. However, the complexity of Go is beyond the agent's capability. The agent can only win odd games but lost many. Consequently, Alpha trusts the agent less at time t_1 and, correspondingly, the bar becomes shorter in Fig. 3b.

The Human-System dimension at time t_1 becomes decisive in Alpha's trust, which is possibly due to three reasons. Firstly, Alpha may take the number of games that the agent has won as the key criterion for the agent's trustworthiness. Secondly, the VR device used by Alpha is presumed to perform stably as an information system and, therefore, the perceived uncertainty from the VR device decreases over time, resulting in the weakened impact of the Human-Technology dimension on trust. Lastly, the Interpersonal dimension may be less concerning to Alpha since the system performance appears to be the most prominent issue, and the cartoonish anthropomorphic features are likely insignificant to a thirty-year-old person's decisions on delegation. Nevertheless, the Interpersonal dimension still plays a role at the beginning ($t = t_0$) since the first impression of the agent can be an important base for initial trust.

By time t_2, the agent may have lost many games and, thus, Alpha starts to distrust the agent, which is illustrated in Fig. 3c as the bar below the plane.

The bars at $S2D1$ illustrate the temporal dynamics of Alpha's trust in the agent when playing another game – Gomoku. The two activities, i.e., playing Go and Gomoku, are independent from each other. Given the lower complexity of Gomoku compared with Go, Alpha is likely to have a higher level of initial trust in the agent. Between t_0 and t_1, the agent is fully competent in playing Gomoku. Consequently, in contrast to the Go situation, Alpha's trust has increased by time t_1, which is illustrated by the bar at $S2D1$ in Fig. 3b. At time t_1, the Human-System dimension continues to play a significant role in Alpha's trust since the agent's performance may still predominate in Alpha's concerns. Nonetheless, the system performance issue is less perturbing and, consequently, the Interpersonal dimension may have a relatively more pronounced impact than that in the Go situation. With the competence of the agent, Alpha's trust continues to increase as the interaction proceeds, and the bar becomes even higher at time t_2 as illustrated in Fig. 3c.

The bars at $S2D2$ illustrate the trust dynamics of a different user –Beta, a five-year-old child– in the agent in the context of playing Gomoku. Beta forms trust in a similar fashion as Alpha. However, the Interpersonal dimension may have a pronounced impact on Beta

[2] Gomoku is also called five-in-a-row game, which uses the board and pieces of Go but follows a different and much simpler set of rules.

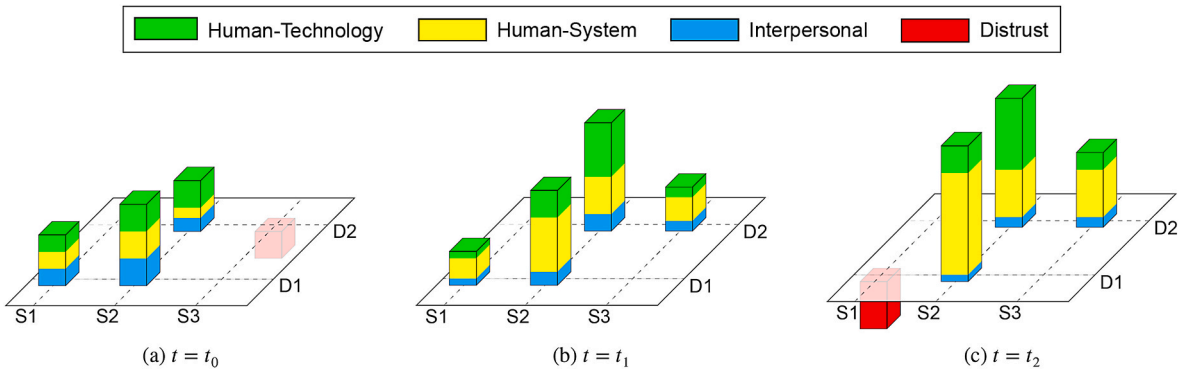

Fig. 3. Trust dynamics over time. The subfigures illustrate the changing influence of the different trust dimensions on trust in the same agent during an interaction at time t_0, t_1, and t_2. The colored bars forming the positive trust landscape show the relative dimensional influence for dispositions $D1$ and $D2$ in situations $S1$, $S2$, and $S3$. The red, negative bars represent distrust. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

because a child's trust may simply depend on it liking or disliking the agent's cartoonish cat appearance.

The bars at $S3D2$ illustrate the dynamics of Beta's trust in the agent in the context of playing a niche game, which Beta may have already skilled. Thus, in the beginning, Beta may distrust the agent out of self-confidence and prefers to play the game personally. At time t_1, Beta may have found that the agent is proficient at playing this game and, therefore, starts to show an inverted trusting attitude toward the agent. Later, Beta may have found him/herself outperformed by the agent, which results in the further increase of Beta's trust as illustrated in Fig. 3c.

4.4. Model validation

The suggested set of dimensions constitutes an abstract construct for assessing trust formation in IAA when multiple factors are involved or considered. It provides a theoretical basis for inspecting and designing trustworthy IAA; therefore, validating the model and its underlying dimensions is important to substantiate the model and prove its practicality.

Due to the inherent complexity and multiple-factor analysis, extensive validation is time-consuming and requires a series of different, dedicated studies. We are already applying and assessing the model in various contexts, such as critical delegation decisions to IAA, and will discuss further potential validation in the following.

Taking a constructive approach and applying principal component analysis (PCA) to the intermediate factors, for instance, can provide valuable evidence justifying the three dimensions of the model. As a common technique for reducing the number of variables in a data set, PCA selects and combines them into new variables that describe the data set more precisely based on their relevance. This process resembles the construction of the model itself, where the intermediate factors are classified into dimensions based on their impact and relevance. The suggested set of dimensions can be considered valid if the PCA results in a similar number of significant components and their correlations with the various intermediate factors are consistent with the model.

Another approach to model validation is through meta-cognition, employing well-established assessment methods such as the *Trust in Technology* or *Trust in Automation* questionnaires (Jian et al., 2000; Mcknight et al., 2011). As discussed earlier in this section, the different dimensions in the model are directly related to trust in technology (Human-Technology), in automation and autonomy (Human-System), as well as their combination (Interpersonal). Consequently, the metrics intended for these categories may equally apply to validating the model dimensions, which are discretized to then compare prediction and truth. Since such metrics might be inaccurate or erroneous, the obtained results must be statistically significant and ideally based on a large sample size. Self-reports ideally are combined with behavior-based measurement, and we are developing assessment strategies involving both.

5. Open issues

Trust has been the subject of research for many decades, resulting in many significant findings. Yet, despite the substantial advances in trust research, there still is no methodical approach to designing trustworthy IAA in VR and many other kinds of systems. A major issue preventing practical, systematic guidelines is the difficulty of integrating and unifying experimental trust research. Most studies were conducted, e.g., under different contexts, different methodologies, making it even harder to integrate them. Thus, we identify an initial set of three open yet fundamental issues concerning research transferability, trust measurement, and representations of self to facilitate future studies. These issues –if considered in the experimental design– can avoid common limitations of user-IAA trust research specific to VR settings. They will be individually discussed in Section 5.1, Section 5.2, and Section 5.3.

5.1. Research transferability

One major problem impeding the current research is the lack of trust-related studies on IAA specific to VR. While there is extensive literature on trust in reality, directly applying their findings into virtual environments is inappropriate. Novel VR devices can create highly realistic and immersive virtual environments, blurring the boundaries between reality and virtual realities. This similarity seemingly allows us to transfer real-world studies directly into VR.

Evidence shows the transferability between the studies in reality and in virtual environments mediated by less immersive VR devices such as flat screens. For example, females are perceived as more trustworthy than males in reality (Buchan et al., 2008), which is consistent with the finding of female avatars being perceived as more trustworthy than male avatars in virtual environments (Surprenant, 2012). Similar effects are also found in the virtual environments mediated by more immersive VR devices (Garau et al., 2005). This agrees with a major opinion in the recent literature, which states that people react to the same stimuli in VR similarly to how they would in reality (Martens et al., 2019). However, there is also evidence showing that people may react differently in VR (Gallup et al., 2019), which reduces the transferability of the research.

The opposite direction of this transferability (i.e., from VR to reality) is also relevant and has been recently discussed by sociologists and psychologists as exemplified in (Pan & Hamilton, 2018), which points out that VR technology is particularly useful in sociological and psychological experiments due to three advantages: absolute experimental control, reproducibility, and a higher level of ecological validity.

Measuring the transferability of a study demands identifying the

factors that are decisive in its transferability. Thus, an important research question can be formulated as

Question 1. What are the factors that determine the transferability of a trust-related study across environments with different levels of immersion?

5.2. Trust metrics

Trust assessment is fundamental to trust-related experiments, and there exist many trust metrics in the literature. The most commonly used trust metric is questionnaires, both custom (Philip et al., 2020) or standardized as in the *Trust in Automation Questionnaire* (Jian et al., 2000), the *Trust in Technology Questionnaire* (Mcknight et al., 2011), or the *Interpersonal Trust Scale* (Rotter, 1967). Verbal interviews are sometimes conducted as a complementary or alternative trust metric to obtain detailed information. While questionnaires and interviews are based on and elicit trustors' opinions, some other trust metrics focus on trustors' behaviors. For instance, in the Trust Game, trust in the opponent can be measured by the money that participants send to the other side (George et al., 2018). Additionally, trust can be measured via physiological signals, such as oxytocin and placebo (De Visser et al., 2017).

However, for two reasons, the trust metrics mentioned above may need reconsideration when measuring trust within virtual environments. Firstly, some trust metrics are conventionally conducted in reality, such as questionnaires. This requires subjects to remove VR devices to initiate the measurement. The context change from VR to reality may cause the *break in presence*, which may lead to biased results. Reimplementing these questionnaire-based metrics in VR may help alleviate the break in presence but demands careful design (Putze et al., 2020). Secondly, behavior-based trust metrics are worth exploring since VR allows designers to realize what would be unrealizable in reality. For instance, subjects' behaviors in a virtual maze can be used as an indicator of trust (Hale et al., 2018). Additionally, the commonly embedded sensors (e.g., eye-gaze detectors, body-tracking devices, VR gloves, and VR treadmills) in VR devices can capture various physiological data, facilitating the analysis of trusting behaviors.

5.3. Trust and representations of self

In reality, individuals typically have a single body over their lifespan. The ability to swap bodies is so far still a fiction in reality, whereas VR allows people to embody different representations of themselves, which may influence the process of establishing trust within VR.

The representation of an individual in VR has been empirically demonstrated to impact the individual's perception, cognition, and behaviors. For instance, white people who embody a black skin virtual avatar in VR have shown sustainable descend in their implicit racial bias (Banakou et al., 2016). Individuals with their avatar's arms textured in red are found to have a lower pain threshold toward heat, for instance compared to blue arms (Martini et al., 2013). Adults have expressed strong body ownership when embodied in a four-year-old child avatar or an avatar resized to the height of a four-year-old (Banakou et al., 2013).

Likewise, we assume that an individual's representation may also influence the individual's trust in IAA. However, to our best knowledge, only few studies have focused on the correlation between trust and representations of self in augmented reality and mixed reality (Jo et al., 2017; Pan & Steed, 2017). There is a need for further research at the intersection of representation, trust, and VR, which leads us to formulate the following research question:

Question 2. What is the influence of users' representations of self in VR on their trust in IAA?

Answering the question above will shed light on another issue that hinders identifying universally trustworthy designs. Comparing to the considerable variations of dispositions and situations, most experiments on trust only recruit a relatively small number of participants, varying from two-digit numbers to small three-digit numbers. We recognize the substantial difficulty in conducting experiments with thousands of demographically different participants. Nonetheless, whether such a limited sample size and variation can lead to generalizable findings remains a question. The concern is also reflected in (Balfe et al., 2018), which points out that trust studies conducted in lab settings may hardly apply to real-world scenarios.

There are mainly two approaches to improve on this issue: increasing sample size and variation or considering the sample-size issue in the analysis of the results. Based on the discussion above, we argue that assigning participants different avatars in VR is a potential alternative.

Question 3. Can different representations of self be employed to augment or diversify samples in the experiments regarding trust within VR?

If this method is justified, research that aims for generalizable findings can be approached by embodying participants in different personas. Conversely, it is also possible to only focus on specific demographics with different participants embodied in the same avatar.

6. Conclusion

Hybrid human-IAA societies are already a reality also outside of a VR context. Trust plays a crucial role in improving their performance and prosperity. User interaction and collaboration with IAA will intensify as technology evolves, attracting cross-disciplinary research to which, for instance, the increasing number of conferences at the intersection of AI and VR bears witness.

For this reason, this article proposed a trust model comprising a Human-Technology, a Human-System, and an Interpersonal dimension. We discussed each of these dimensions in detail to clarify the conceptual differences and identified various factors that influence trust from a broad spectrum of literature. We further elucidated the various interrelations between the different trust dimensions and their dynamic impact on users' trust in IAA. This classification helps to describe intelligent autonomous agents in VR and to comparatively assess their characteristics, acceptance, and trustworthiness. Finally, we discussed open issues regarding research transferability, trust metrics, and representation, which are essential to future user-IAA trust research specific to VR contexts.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–18).

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Adams, D., Bah, A., Barwulor, C., Musaby, N., Pitkin, K., & Redmiles, E. M. (2018). Ethics emerging: The story of privacy and security perceptions in virtual reality. In *Proceedings of the 14th Symposium on useable privacy and security ({SOUPS})* (pp. 427–442).

Aldowah, H., Rehman, S. U., & Umar, I. (2021). Trust in IoT systems: A vision on the current issues, challenges, and recommended solutions. *Advances on Smart and Soft Computing*, 329–339.

Amir, D., & Amir, O. (2018). Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International conference on autonomous agents and MultiAgent systems* (pp. 1168–1176).

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable

artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.
Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science, 16*(10), 814–819.
Balfe, N., Sharples, S., & Wilson, J. R. (2018). Understanding is key: An analysis of factors pertaining to trust in a real-world automation system. *Human Factors, 60*(4), 477–495.
Banakou, D., Groten, R., & Slater, M. (2013). Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *Proceedings of the National Academy of Sciences, 110*(31), 12846–12851.
Banakou, D., Hanumanthu, P. D., & Slater, M. (2016). Virtual embodiment of white people in a black virtual body leads to a sustained reduction in their implicit racial bias. *Frontiers in Human Neuroscience, 10*, 601.
Baylor, A. L. (2009). Promoting motivation with virtual agents and avatars: Role of visual presence and appearance. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), 3559–3565.
Blank, G., & Dutton, W. H. (2012). Age and trust in the Internet: The centrality of experience and attitudes toward technology in Britain. *Social Science Computer Review, 30*(2), 135–151.
Blumberg, B., & Galyean, T. (1997). Multi-level control for animated autonomous agents: Do the right thing … oh, not that. In R. Trappl, & P. Petta (Eds.), *Creating personalities for synthetic actors: Towards autonomous personality agents* (pp. 74–82). Berlin, Heidelberg: Springer Berlin Heidelberg.
Bowman, D. A., Johnson, D. B., & Hodges, L. F. (2001). Testbed evaluation of virtual environment interaction techniques. *Presence, 10*(1), 75–95.
Brundage, M. (2015). Taking superintelligence seriously: Superintelligence: Paths, dangers, strategies by Nick Bostrom (Oxford University Press, 2014). *Futures, 72*, 32–35.
Bruun, E. P., & Duka, A. (2018). Artificial intelligence, jobs and the future of work: Racing with the machines. *Basic Income Studies, 13*(2), 1–15.
Buchan, N. R., Croson, R. T., & Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the investment game. *Journal of Economic Behavior & Organization, 68*(3–4), 466–476.
Cai, H., & Lin, Y. (2012). Coordinating cognitive assistance with cognitive engagement control approaches in human–machine collaboration. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 42*(2), 286–294.
Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review, 63*, 277–293.
Cascio, C. J., Moore, D., & McGlone, F. (2019). Social touch and human development. *Developmental cognitive neuroscience, 35*, 5–11.
Castle, E., Eisenberger, N. I., Seeman, T. E., Moons, W. G., Boggero, I. A., Grinblatt, M. S., & Taylor, S. E. (2012). Neural and behavioral bases of age differences in perceptions of trust. *Proceedings of the National Academy of Sciences, 109*(51), 20848–20852.
Chan, L., Morgan, I., Simon, H., Alshabanat, F., Ober, D., Gentry, J., Min, D., & Cao, R. (2019). Survey of AI in cybersecurity for information technology management. In *Proceedings of 2019 IEEE technology & engineering management conference (TEMSCON)* (pp. 1–8).
Chiou, E. K., & Lee, J. D. (2021). *Trusting automation: Designing for responsivity and resilience. Human factors*. p. 00187208211009995.
Cook, K. S., Yamagishi, T., Cheshire, C., Cooper, R., Matsuda, M., & Mashima, R. (2005). Trust building via risk taking: A cross-societal experiment. *Social Psychology Quarterly, 68*(2), 121–142.
De Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R., Parasuraman, 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.
De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., 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.
Dennett, D. C. (2008). *Kinds of minds: Toward an understanding of consciousness*.
Dennett, D. (2009). Intentional systems theory. In *The Oxford handbook of philosophy of mind* (pp. 339–350).
Deutsch, M. (1973). *The resolution of conflict: Constructive and destructive processes*.
Din, I. U., Bano, A., Awan, K. A., Almogren, A., Altameem, A., & Guizani, M. (2021). Light Trust: Lightweight trust management for edge devices in industrial Internet of Things. *IEEE Internet of Things J., 1–1*.
Dong, W., Yang, T., Liao, H., & Meng, L. (2020). How does map use differ in virtual reality and desktop-based environments? *Int. J. Digital Earth, 13*(12), 1484–1503.
Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: The influence of emotion on trust. *Journal of Personality and Social Psychology, 88*(5), 736.
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.
Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C., et al. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence, 3*(7), 566–571.
Felnhofer, A., Heinzle, A.k., & Kothgassner, O. D. (2013). Game experience and behavior in young women: A comparison of interface technologies. In *Proceedings of the 11th usability day on natural user interfaces* (pp. 1–6).
Fogg, B., & Nass, C. (1997). How users reciprocate to computers: An experiment that demonstrates behavior change. In *Proceedings of the 1997 CHI extended abstracts on human factors in computing systems* (pp. 331–332).
Fortino, G., Fotia, L., Messina, F., Rosaci, D., & Sarné, G. M. (2020). Trust and reputation in the Internet of Things: State-of-the-Art and research challenges. *IEEE Access, 8*, 60117–60125.
Fortino, G., Fotia, L., Messina, F., Rosaci, D., & Sarné, G. M. (2021). A blockchain-based group formation strategy for optimizing the social reputation capital of an IoT scenario. *Simulation Modelling Practice and Theory, 108*, 102261.
Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Proceedings of the 1996 International workshop on agent theories, architectures, and languages (ATAL)* (pp. 21–35).
Friedman, B., Khan, P. H., Jr., & Howe, D. C. (2000). Trust online. *Communications of the ACM, 43*(12), 34–40.
Gallimore, D., Lyons, J. B., Vo, T., Mahoney, S., & Wynne, K. T. (2019). Trusting robocop: Gender-based effects on trust of an autonomous robot. *Frontiers in Psychology, 10*, 482.
Gallup, A. C., Vasilyev, D., Anderson, N., & Kingstone, A. (2019). Contagious Yawning in virtual reality is affected by actual, but not simulated, social presence. *Scientific Reports, 9*(1), 1–10.
Garau, M., Slater, M., Pertaub, D. P., & Razzaque, S. (2005). The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators and Virtual Environments, 14*(1), 104–116.
García-Magariño, I., Muttukrishnan, R., & Lloret, J. (2019). Human-centric AI for trustworthy IoT systems with explainable multilayer perceptrons. *IEEE Access, 7*, 125562–125574.
George, C., Eiband, M., Hufnagel, M., & Hussmann, H. (2018). Trusting strangers in immersive virtual reality. In *Proceedings of the 2018 International conference on intelligent user interfaces companion* (pp. 1–2). (IUI).
Ghahramani, M., Qiao, Y., Zhou, M., Hagan, A. O., & Sweeney, J. (2020). AI-based modeling and data-driven evaluation for smart manufacturing processes. *IEEE/CAA Journal of Automatica Sinica, 7*(4), 1026–1037.
Gong, L., & Nass, C. (2007). When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human Communication Research, 33*(2), 163–193.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*, 2672–2680.
Guerrero, L. K., & Floyd, K. (2006). *Nonverbal communication in close relationships*.
Hale, J., & Antonia, F. D. C. (2016). Testing the relationship between mimicry, trust and rapport in virtual reality conversations. *Scientific Reports, 6*(1), 1–11.
Hale, J., Payne, M. E., Taylor, K. M., Paoletti, & Davide sand De C Hamilton, A. F. (2018). The virtual maze: A behavioural tool for measuring trust. *Quarterly Journal of Experimental Psychology, 71*(4), 989–1008.
Hanna, N., & Richards, D. (2018). The impact of multimodal communication on a shared mental model, trust, and commitment in human–intelligent virtual agent teams. *Multimodal Technologies and Interaction, 2*(3), 48.
Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.
Hriez, S. F., Almajali, S., & Ayyash, M. (2021). Trust models in IoT-enabled WSN: A review. In *International conference on data science, E-learning and information systems 2021* (pp. 153–159). New York, NY, USA: Association for Computing Machinery.
Huerta, E., Glandon, T., & Petrides, Y. (2012). Framing, decision-aid systems, and culture: Exploring influences on fraud investigations. *International Journal of Accounting Information Systems, 13*(4), 316–333.
Huisman, G. (2017). Social touch technology: A survey of haptic technology for social touch. *IEEE Transactions on Haptics, 10*(3), 391–408.
Ide, S. (1982). Japanese sociolinguistics politeness and women's language. *Lingua, 57*(2–4), 357–385.
Imam, I. F., & Kodratoff, Y. (1997). Intelligent adaptive agents: A highlight of the field and the AAAI-96 workshop. *AI Magazine, 18*(3), 75, 75.
Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill, 4*(2), e14.
Javaid, M., & Haleem, A. (2020). Virtual reality applications toward medical field. *Clinical Epidemiology and Global Health, 8*(2), 600–605.
Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined Scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71.
Johnson, J. D., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Type of automation failure: The effects on trust and reliance in automation. In *Proceedings of the 2004 human factors and ergonomics society annual meeting* (Vol. 48, pp. 2163–2167).
Jo, D., Kim, K. H., & Kim, G. J. (2017). Effects of avatar and background types on users' Co-presence and trust for mixed reality-based teleconference systems. In *Proceedings the 2017 conference on computer animation and social agents (CASA)* (pp. 27–36).
Kravari, K., & Bassiliades, N. (2019). StoRM: A social agent-based trust model for the Internet of Things adopting microservice architecture. *Simulation Modelling Practice and Theory, 94*, 286–302.
Krumhuber, E., Manstead, A. S., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion, 7*(4), 730.
Lample, G., & Chaplot, D. S. (2017). Playing FPS games with deep reinforcement learning. In *Proceedings of the thirty-first AAAI conference on artificial intelligence* (pp. 2140–2146). San Francisco, California, USA: AAAI'17, AAAI Press.
Lee, E. J. (2010). What triggers social responses to flattering computers? Experimental tests of anthropomorphism and mindlessness explanations. *Communication Research, 37*(2), 191–214.
Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270.
Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.
Lee, Y. H., Xiao, M., & Wells, R. H. (2018). The effects of avatars' age on older adults' self-disclosure and trust. *Cyberpsychology, Behavior, and Social Networking, 21*(3), 173–178.

Legacy, C., Ashmore, D., Scheurer, J., Stone, J., & Curtis, C. (2019). Planning the driverless city. *Transport Reviews, 39*(1), 84–102.

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Signed networks in social media. In *Proceedings the SIGCHI conference on human factors in computing systems (CHI'10)* (pp. 1361–1370).

Li, Y., Hao, Z., & Lei, H. (2016). Survey of convolutional neural network. *Journal of Computer Applications, 36*(9), 2508–2515.

Lippert, S. K. (2002). *An exploratory study into the relevance of trust in the context of information systems technology*. Ph.D. thesis.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650.

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.

Luck, M., d'Inverno, M., et al. (1995). A formal framework for agency and autonomy. *ICMAS, 95*, 254–260.

Lugano, G. (2017). Virtual assistants and self-driving cars. In *Proceedings of the 2017 International conference on ITS telecommunications (ITST)* (pp. 1–5).

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors, 48*(2), 241–256.

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.

Marsh, S. P. (1994). *Formalising trust as a computational concept*. Ph.D. thesis.

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

Martens, M. A., Antley, A., Freeman, D., Slater, M., Harrison, P. J., & Tunbridge, E. M. (2019). It feels real: Physiological responses to a stressful virtual reality environment and its impact on working memory. *Journal of Psychopharmacology, 33*(10), 1264–1273.

Martini, M., Pérez Marcos, D., & Sanchez-Vives, M. V. (2013). What color is my arm? Changes in skin color of an embodied virtual arm modulates pain threshold. *Frontiers in Human Neuroscience, 7*, 438.

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

McDonnell, R., Breidt, M., & Bülthoff, H. H. (2012). Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG), 31*(4), 1–11.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89–94.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS), 2*(2), 1–25.

McKnight, D. H., & Chervany, N. L. (1996). *The Meanings of trust*. Tech. rep.

de Melo, C. M., Gratch, J., & Carnevale, P. J. (2014). Humans versus computers: Impact of emotion expressions on people's decision making. *IEEE Transactions on Affective Computing, 6*(2), 127–136.

Mishra, P. (2006). Affective feedback from computers and its effect on perceived ability and affect: A test of the computers as social actor hypothesis. *Journal of Educational Multimedia and Hypermedia, 15*(1), 107–131.

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research, 26*(4), 323–339.

Moon, Y., & Nass, C. (1996). How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication Research, 23*, 651–674.

Mori, M. (1970). Bukimi no tani [the Uncanny Valley]. *Energy, 7*, 33–35.

Morra, L., Lamberti, F., Pratticó, F. G., La Rosa, S., & Montuschi, P. (2019). Building trust in autonomous vehicles: Role of virtual reality driving simulators in HMI design. *IEEE Transactions on Vehicular Technology, 68*(10), 9438–9450.

Moruzzi, C. (2017). Creative AI: Music composition programs as an extension of the composer's mind. In *Proceedings of the 2017 conference on philosophy and theory of artificial intelligence (PT-AI)* (pp. 69–72).

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905–1922.

Myers, C. D., & Tingley, D. (2016). The influence of emotion on trust. *Political Analysis, 24*(4), 492–500.

Nakajima, K., & Niitsuma, M. (2020). Effects of space and scenery on virtual pet-assisted activity. In *Proceedings of the 8th International conference on human-agent interaction* (pp. 105–111).

Narang, S., Best, A., & Manocha, D. (2019). Inferring user intent using Bayesian theory of mind in shared avatar-agent virtual environments. *IEEE Transactions on Visualization and Computer Graphics, 25*(5), 2113–2122.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology, 27*(10), 864–876.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the 1994 CHI conference on human factors in computing systems* (pp. 72–78).

Nowak, K. L., & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 12*(5), 481–494.

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.

Pan, X., & Hamilton, A. F.d. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology, 109*(3), 395–417.

Pan, Y., & Steed, A. (2017). The impact of self-avatars on trust and collaboration in shared virtual environments. *PLoS One, 12*(12), Article e0189078.

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

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). GauGAN: Semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH 2019 real-time live* (p. 1).

Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology, 110*(6), 785.

Perkis, A., Timmerer, C., et al. (2020). *QUALINET white paper on definitions of immersive media experience*. (IMEx).

Peukert, C., Pfeiffer, J., Meißner, M., Pfeiffer, T., & Weinhardt, C. (2019). Shopping in virtual reality stores: The influence of immersion on system Adoption. *Journal of Management Information Systems, 36*(3), 755–788.

Philip, P., Dupuy, L., Auriacombe, M., Serre, F., de Sevin, E., Sauteraud, A., & Micoulaud-Franchi, J. A. (2020). Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *NPJ Digital Medicine, 3*(1), 1–7.

Picard, R. W. (2000). *Affective computing*.

Pietroszek, K., & Lee, N. (2019). *Virtual hand metaphor in virtual reality*.

Pliatsios, D., Sarigiannidis, P., Efstathopoulos, G., Sarigiannidis, A., & Tsiakalos, A. (2020). Trust management in smart grid: A markov trust model. In *Proceedings of the 2020 International conference on modern circuits and systems technologies* (pp. 1–4). (MOCAST).

Price, S., Jewitt, C., & Yiannoutsou, N. (2021). Conceptualising touch in VR. *Virtual Reality*, 1–15.

Putze, S., Alexandrovsky, D., Putze, F., Höffner, S., Smeddinck, J. D., & Malaka, R. (2020). Breaking the experience: Effects of questionnaires in VR user studies. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–15).

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*(1), 95.

Rickel, J., & Johnson, W. L. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence, 13*(4–5), 343–382.

Riek, L. D. (Jul 2012). Wizard of oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction, 1*(1), 119–136.

Righi, V., Sayago, S., & Blat, J. (Dec 2017). When we talk about older people in HCI, who are we talking about? Towards A 'turn to community' in the design of technologies for A growing ageing population. *International Journal of Human-Computer Studies, 108*, 15–31.

Roselyn Lee, J. E., Nass, C., Brave, S. B., Morishima, Y., Nakajima, H., & Yamada, R. (2007). The case for caring colearners: The effects of a computer-mediated colearner agent on trust and learning. *Journal of Communication, 57*(2), 183–204.

Ross, J. (2008). *Moderators of trust and reliance across multiple decision aids*. Ph.D. thesis.

Rotter, J. B. (1967). A new Scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651–665.

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.

Sagnier, C., Loup-Escande, E., Lourdeaux, D., Thouvenin, I., & Valléry, G. (2020). User acceptance of virtual reality: An extended technology acceptance model. *International Journal of Human-Computer Interaction, 36*(11), 993–1007.

Sajeev, M., Cohen, J., Wakefield, C. E., Fardell, J. E., & Cohn, R. J. (2017). Decision aid for nutrition support in pediatric oncology: A pilot study. *Journal of Parenteral and Enteral Nutrition, 41*(8), 1336–1347.

Salanitri, D. (2018). *Trust in virtual reality*. Ph.D. thesis.

Salanitri, D., Hare, C., Borsci, S., Lawson, G., Sharples, S., & Waterfield, B. (2015). Relationship between trust and usability in virtual environments: An ongoing study. In *Proceedings of the 2015 International conference on human-computer interaction (CHII)* (pp. 49–59).

Salanitri, D., Lawson, G., & Waterfield, B. (2016). The relationship between presence and trust in virtual reality. In *Proceedings of the 2016 European conference on cognitive ergonomics (ECCE)* (pp. 1–4).

Salanitri, D., Lawson, G., & Waterfield, B. (2020). Enhancing trust in virtual reality systems. In *New perspectives on virtual and augmented reality: Finding new ways to teach in a transformed learning environment* (pp. 132–146).

Santos, B. S., Dias, P., Pimentel, A., Baggerman, J. W., Ferreira, C., Silva, S., & Madeira, J. (2009). Head-mounted display versus desktop for 3D navigation in virtual reality: A user study. *Multimedia Tools and Applications, 41*(1), 161–181.

Scanzoni, J. (1979). *Social exchange and behavioral interdependence*.

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

Schuemie, M. J., Van Der Straaten, P., Krijn, M., & Van Der Mast, C. A. (2001). Research on presence in virtual reality: A survey. *CyberPsychology and Behavior, 4*(2), 183–201.

Seckler, M., Heinz, S., Forde, S., Tuch, A. N., & Opwis, K. (2015). Trust and distrust on the web: User experiences and website characteristics. *Computers in Human Behavior, 45*, 39–50.

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence, 60*(1), 51–92.

Song, X., Yang, S., Huang, Z., & Huang, T. (2019). The application of artificial intelligence in electronic commerce. *Journal of Physics: Conference Series*, 32030, 1302.

Stein, J. P., & Ohler, P. (2017). Venturing into the Uncanny Valley of mind — the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition, 160*, 43–50.

Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., & Speranza, N. (2010). Accounting for the human in cyberspace: Effects of mood on trust in automation. In *Proceedings of 2010 International Symposium on collaborative technologies and systems (CTS)* (pp. 180–187).

Stout, N., Dennis, A. R., & Wells, T. M. (2014). The buck stops there: The impact of perceived accountability and control on the intention to delegate to software agents. *AIS Transactions on Human-Computer Interaction, 6*(1), 1–15.

Surprenant, A. M. (2012). *Measuring trust in virtual worlds: Avatar-mediated self-disclosure*. Ph.D. thesis.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied, 17*(2), 71.

Takayama, L. (2009). Making sense of agentic objects and teleoperation: In-the-Moment and reflective perspectives. In *2009 4th ACM/IEEE International conference on human-robot interaction (HRI)* (pp. 239–240). IEEE.

Teoh, E. R., & Kidd, D. G. (2017). Rage against the machine? Google's self-driving cars versus human drivers. *Journal of Safety Research, 63*, 57–60.

Thielsch, M. T., Meeßen, S. M., & Hertel, G. (2018). Trust and distrust in information systems at the workplace. *PeerJ, 6*, e5483.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519–545.

Torre, I., Carrigan, E., McDonnell, R., Domijan, K., McCabe, K., & Harte, N. (2019). The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction. In *Proceedings of the 2019 motion, interaction and games conference* (pp. 1–6). (MIG).

Tulshan, A. S., & Dhage, S. N. (2018). Survey on virtual assistant: Google assistant, siri, cortana, Alexa. In *The 2018 International Symposium on signal processing and intelligent recognition systems (SIRS)* (pp. 190–201).

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature, 575*(7782), 350–354.

Wakefield, R. L., Stocks, M. H., & Wilder, W. M. (2004). The role of web site characteristics in initial trust formation. *Journal of Computer Information Systems, 45*(1), 94–103.

de Weerd, H., Verbrugge, R., & Verheij, B. (2017). Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems, 31*(2), 250–287.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 2019 ACM International conference on intelligent virtual agents* (pp. 7–9). IVA.

Wooldridge, M. (1999). Intelligent Agents. *Multiagent Systems, 6*, 27–77.

Yee, N., & Bailenson, J. (Jul 2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research, 33*(3), 271–290.

Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors, 43*(3), 355–365.

Yuksel, B. F., Collisson, P., & Czerwinski, M. (2017). Brains or beauty: How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT), 17*(1), 1–20.

Yu, H., Shen, Z., Leung, C., Miao, C., & Lesser, V. R. (2013). A survey of multi-agent trust management systems. *IEEE Access, 1*, 35–50.

Zaltman, G., & Moorman, C. (1988). The importance of personal trust in the use of research. *Journal of Advertising Research, 28*(5), 16–24.