Research Article

Alexander M. Aroyo, Jan de Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz*, Henrik Sætra, Mads Solberg, and Aurelia Tamò-Larrieux

Overtrusting robots: Setting a research agenda to mitigate overtrust in automation

https://doi.org/10.1515/pjbr-2021-0029
received May 21, 2021; accepted August 28, 2021

Abstract: There is increasing attention given to the concept of trustworthiness for artificial intelligence and robotics. However, trust is highly context-dependent, varies among cultures, and requires reflection on others' trustworthiness, appraising whether there is enough evidence to conclude that these agents deserve to be trusted. Moreover, little research exists on what happens when too much trust is placed in robots and autonomous systems. Conceptual clarity and a shared framework for approaching overtrust are missing. In this contribution, we offer an overview of pressing topics in the context of overtrust and robots and autonomous systems. Our review mobilizes insights solicited from in-depth conversations from a multidisciplinary workshop on the subject of trust in human–robot interaction (HRI), held at a leading robotics conference in 2020. A broad range of participants brought in their expertise, allowing the formulation of a forward-looking research agenda on overtrust and automation biases in robotics and autonomous systems. Key points include the need for multidisciplinary understandings that are situated in an eco-system perspective, the consideration of adjacent concepts such as deception and anthropomorphization, a connection to ongoing legal discussions through the topic of liability, and a socially embedded understanding of overtrust in education and literacy matters. The article integrates diverse literature and provides a ground for common understanding for overtrust in the context of HRI.

Keywords: trust, overtrust, robots, social robots, deception, anthropomorphization liability, education

*** Corresponding author: Christoph Lutz,** Nordic Centre for Internet and Society, BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway, e-mail: christoph.lutz@bi.no
Alexander M. Aroyo: SIRRL, University of Waterloo, Waterloo, Canada, e-mail: alexander.aroyo@uwaterloo.ca
Jan de Bruyne: KU Leuven Centre for IT & IP Law (CiTiP), KU Leuven, Leuven, Belgium, e-mail: jan.debruyne@kuleuven.be
Orian Dheu: KU Leuven Centre for IT & IP Law (CiTiP), KU Leuven, Leuven, Belgium, e-mail: orian.dheu@kuleuven.be
Eduard Fosch-Villaronga: eLaw Center for Law and Digital Technologies, Leiden University, Leiden, The Netherlands, e-mail: e.fosch.villaronga@law.leidenuniv.nl
Aleksei Gudkov: Faculty of Law: School for Theory of Law and Cross-sectoral Legal Disciplines, National Research University Higher School of Economics, Moscow, Russia, e-mail: avgudkov@hse.ru
Holly Hoch: FAA-HSG, University of St. Gallen, St. Gallen, Switzerland, e-mail: holly.hoch@student.unisg.ch
Steve Jones: Department of Communication, University of Illinois Chicago, Chicago, United States, e-mail: sjones@uic.edu
Henrik Sætra: Department of Computer Science and Communication, Østfold University College, Østfold, Norway, e-mail: henrik.satra@hiof.no
Mads Solberg: Department of Health Sciences in Ålesund, Norwegian University of Science and Technology, Trondheim, Norway, e-mail: mads.solberg@ntnu.no
Aurelia Tamò-Larrieux: FAA-HSG, University of St. Gallen, St. Gallen, Switzerland, e-mail: aurelia.tamo@unisg.ch
ORCID: Alexander M. Aroyo 0000-0003-2445-4026;
Jan de Bruyne 0000-0002-9762-3567;
Orian Dheu 0000-0001-6454-5251;
Eduard Fosch-Villaronga 0000-0002-8325-5871;
Aleksei Gudkov 0000-0002-3789-3813;
Holly Hoch 0000-0002-2517-9487;
Steve Jones 0000-0003-1198-9849;
Christoph Lutz 0000-0003-4389-6006;
Henrik Sætra 0000-0002-7558-6451;
Aurelia Tamò-Larrieux 0000-0003-3404-7643

1 Introduction

We live in a time of increasing reliance on algorithmic systems for a multitude of activities. Simple activities, such as using an app to find the nearest shared bike, as well as more complex tasks, like voting or getting a job, are often directly affected by decision-making processes carried out by algorithms. When decision-making agents become embodied, as is the case for social robots, understanding trust and overtrust becomes particularly important as these agents increasingly become parts of natural contexts in which developers have little control of the robot's surroundings. In addition, the increased human-likeness of

Open Access. © 2021 Alexander M. Aroyo *et al.*, published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

certain robots invites humans to trust them in a way distinct from that of, say, a robot arm in a car factory. In the words of the European Commission, "as digital technology becomes an ever more central part of every aspect of people's lives, people should be able to trust it. Trustworthiness is also a prerequisite for its uptake" [1]. In this article, we refer to robots or robotics when we mean embodied agents, while the term autonomous system is used more broadly to include embodied as well as virtual systems that interact with individuals.

In robotics and autonomous system literature, the most common definitions of trust in automation are the following: Lee and See define trust "as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [2]. Meyerson et al. defined trust as "a unique form of collective perception and relating that is capable of managing issues of vulnerability, uncertainty, risk and expectations" [3], while Hancock characterized trust as a reliance on others not to perform actions that are prejudicial to their own well-being [4]. Trust is thus a relational concept, as it is used to describe the trustor's positive expectations of the trustee in uncertain circumstances [5]. Levine and Schweitzer [6] distinguished between what they call benevolence-based trust and integrity-based trust, in which the first regards the goodness and intentions of the trustee and the latter concerns adherence to acceptable ethical principles. Furthermore, trust is highly context-dependent and varies among cultures [7]. It requires reflection on others' trustworthiness, appraising whether there is enough evidence to conclude that these agents deserve to be trusted. The most salient criteria for trustworthiness are truthfulness, lack of exploitation of the dependent party's vulnerabilities, the constructive contribution to expected benefits, and the willingness by the trusted party to be held accountable [8].

Unfortunately, highly automated systems (and especially robots) are at risk of being overtrusted. Lee and See defined overtrust as an insufficient calibration between the system's capabilities and a person's trust [2]. Parasuraman and Riley added that a system can compromise safety and profitability if it has been inappropriately relied upon [9]. Yet, while policymakers in Europe highlight the importance of trust in a (robotically) mediated world [1,10], research on what happens when too much trust is placed in robots and autonomous systems is quite novel [11,12]. Therefore, we call for new research on overtrust not only in the robotics community but also among social scientists, legal scholars, and policymakers.

On 4 September 2020, we held a half-day workshop at the 29th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN).[1] The workshop was held online through Zoom and focused on sharing knowledge about how to understand and address overtrust in robots. It brought together a group of researchers and practitioners from various disciplines, including law, ethics, communication, philosophy, science and technology studies, cognitive anthropology, computer science, and robotics engineering. A call for extended abstracts was circulated on social media (Twitter) and through the personal networks of the organizers. This resulted in seven accepted contributions that were subsequently presented and discussed at the workshop. The topics covered the black box problem/transparency, the notions of overtrust leading to over-reliance, liability risks, deception, the role of robots in human participant research, and a spotlight on the Wizard of Oz (WoZ) technique.

In addition, a leading human–robot interaction (HRI) researcher held a keynote on moral psychology in HRI. Finally, around 15 RO-MAN attendants who did not present an extended abstract joined the workshop and the discussions, leading to an overall number of 25 participants. The workshop participants investigated what happens when trust becomes overtrust, and automation bias, erroneous belief in technological capabilities, seeps in. They analysed scenarios in which trust in automation becomes destructive for individuals, groups of individuals, and society at large; in particular, when overtrust leads to physical (e.g. injuries, damages) and psychological (e.g. manipulation) consequences. The workshop participants were asked during the last interactive session of the workshop to identify the key topics and issues discussed throughout the morning. Each participant was given a chance to respond to the comments of the others. During these discussions, the moderators' task was to distil and re-group common issues that were raised to determine a research agenda that covered each participant's central topics. This brainstorming and synthesis activity aimed to distil critical issues that will (and should) be subject of further research in the coming years concerning overtrust and HRI.

Following up on the workshop, the moderators reached out to the presenters and selected participants to ask if they were interested in writing an agenda paper. Most presenters and approached participants agreed to be involved. To start the joint writing project, the involved authors developed the research agenda identified during the workshop and elaborated further upon those topics. In subsequent months, the writing happened as a collective effort, with each co-author bringing in their expertise on the topic. The paper's goal

1 See http://ro-man2020.unina.it/.

was to synthesize the workshop discussions and expand on them by developing a forward-looking research agenda on overtrust and robots.

In this article, we would like to share this agenda with the larger research community. Below we combine a summary of important discussions during the workshop and identify research gaps.

2 What we know about overtrust and robots

HRI has started to look into overtrust in the context of robots. Wagner et al. provided an accessible overview of the literature and some earlier research [12]. They described overtrust as "a situation in which a person misunderstands the risk associated with an action because the person either underestimates the loss associated with a trust violation; underestimates the chance the robot will make such a mistake; or both" (p. 22). Using the example of paediatric exoskeletons, they showed how children may overtrust the device's capabilities, overlooking their limitations, and how parents may be too emotionally invested to appropriately assess risks.

Broadly speaking the concept of overtrust falls within the topic of misuse of automation, i.e. using automation under false beliefs leading to its inappropriate use [9]. Overtrust is related to automation bias [13], where individuals think that technology is more capable than it really is, and complacency [14], when individuals become less vigilant towards the functioning of a system even though attention should be paid to it. Traditionally, key technologies investigated in research on overtrust have been autopilots and automation in factories [9,14,15]. However, Wagner and colleagues stress the heightened importance of research on the topic with advances in AI and robotics, as an increasing number of systems have become more autonomous in recent years [12].

Even if the literature on overtrust is still emerging, it builds upon a rich literature on trust in automation overall, e.g. refs [2,16–18], and in particular research that analysed the impact on trust when errors are introduced, e.g. refs [2,19,20]. Indeed, the introduction of errors or malfunctioning of systems leads to recalibration of the possible initial positive expectations towards the working of a system (so-called "positivity bias" [18, p. 699]). In other words, the human operator or a system must recalibrate their trust in the system and implement control strategies for it [20]. However, in cases of overtrust, either the errors are not perceived *per se*, or a recalibration process does not occur. This is why the topic of overtrust is, as mentioned, interlinked with complacency and automation bias. In contrast to complacency though (understood as a lack of vigilance over the functioning of a system), overtrust is understood as having a more active meaning, as an individual is actively assuming a greater capability of the system than they should [14].

Even though the concept and definition of overtrust in automation were introduced by the end of the 1990s [2,9], the HRI community has only recently focused on overtrust. Depending on its severity, overtrust can have different consequences. For example, Gaudiello et al. have studied the effects of human conformity following robot suggestions [21]. They found that individuals have a tendency to adapt and change their answers in order to comply with a robot's suggestions. This outcome could be useful for educational purposes, such as when the robot works as a tutor for a child. However, this case also sits at the borderline of overtrust. While the compliance with the robot's suggestion might not be risky, and sometimes even desirable (especially when a robot has more up-to-date or better quality information than the user), it could also create a chain of events whereby trust in a system leads to overtrust.

Salem et al. studied the effects of a robot presenting a faulty behaviour, and how participants would react to the odd requests by the robot [22]. The participants performed most of the tasks requested by the robot even if they were causing property damage or information leakage. The authors concluded that "although the robot's erratic behaviour affected its perceived reliability and trustworthiness, this had no impact on participants' willingness to comply with its instructions, even in the case of unusual requests" [22, p. 7]. Another example of overtrust is provided by Robinette et al., where participants were put into a realistic emergency scenario and still had a strong tendency to follow a robot that earlier expressed faulty behaviour, rather than following the emergency signs [11].

Few researchers in HRI have looked at overtrust from the psychological perspective. Social Engineering (SE) is a psychological manipulation that induces people to take actions that may or may not be in their best interest [23]. These techniques may take advantage of the trust and kindness displayed by humans to conspecifics [24]. Postnikoff and Goldberg started working on the theoretical concept of SE robots [25], while Booth et al. used a robot as a proxy to enter an unauthorized location by piggybacking [26]. Aroyo et al. further worked on a practical application by mapping an already perfected SE framework on a robotic setting where the robot succeeded to acquire personal sensitive information from participants, made them conform to its suggestions, and gamble money [27]. Finally, Aroyo et al. studied the effect of authority

from a highly human-like robot on participants, making them obey morally controversial requests that violate pre-established rules [28].

This overview of the literature shows that overtrusting robots can have adverse consequences. Although trust is essential for HRI, our disposition may effectively generate overtrust that could potentially endanger humans physically and psychologically. The overview also shows that the study of overtrust in robots has mainly been a topic in HRI research and social psychology but has received less attention in other disciplines such as philosophy, the law, sociology, and communication. To provide a broader take on the topic that includes perspectives from the social sciences and humanities, in addition to technical considerations, the participants of the IEEE RO-MAN workshop set out a research agenda summarizing the main discussion points of the workshop and providing an outlook on new inquiry on this research topic. However, it is important to clarify that this research agenda does not aim at exhaustively listing all the topics that could possibly affect the level of trust invested in robots. It is clear that many other factors may be identified and analysed with respect to overtrust. Rather, this overview highlights some of the most pressing discussion points that ought to be further assessed and explored.

3 Overtrust research agenda 2021

Overtrust is inherently linked to trust, but goes beyond a level that is rationally or objectively ideal or suitable. In a sense, we notice a gap or mismatch between what is and what is expected – this gap can be explained by overtrust. As a workshop participant put it, "Intuitively, it [overtrust] seems like a very strange concept as it puts a scale on trusting relationships, but one that extends past the level of optimal or desirable trust, or even sufficient trust." Borenstein et al. state that in the context of robotics, overtrust portrays a situation in which a person accepts risk because the robot appears to, or is expected to, perform a function it cannot, or, in another context, when a person engages in risky behaviours because they believe that the robot will help mitigate the risk [29]. As elaborated above in the literature review (see Section 2), there are still many aspects of overtrust in automation, and overtrust in robotics in particular, that require multidisciplinary research to better distil the relevant factors that contribute to situations which extend the "level of desirable trust." The following research agenda points to the key and, in our opinion, most pressing, topics identified during the IEEE RO-MAN 2020 workshop (see Section 1):

1. The term "overtrust" requires further interdisciplinary conceptualization to make research on the topic more comparable across disciplines and weave in lessons learned from each field into the policymaking discourse;
2. The effects and implications of deception on trust and overtrust must be understood more comprehensively as manipulation by autonomous systems is a key challenge of new technological developments (e.g. artificially intelligent systems, *see* the proposal for AI Act in the EU [30]);
3. The role of anthropomorphization in trusting relationships must be better understood in automation and specifically in the context of embodied agents interacting with individuals (e.g. social robots) to address potential biases and misconceptions formed especially by more vulnerable individuals (e.g. children, elderly, marginalized communities);
4. An analysis of the ecosystem(s) in which overtrust most often occurs should be conducted to identify roles and responsibilities and ensure systems are safe to use (to the whole extent of the word, *see* the dimensions of safety in ref. [31]) and designed in a transparent manner that allows us to allocate accountability [32];
5. Regulatory tools such as liability regulation need to be analysed towards creating appropriate risk-allocation among different stakeholders; and
6. Literacy and educational initiatives must be developed to empower users to identify potentially undesirable trusting relationships with automation.

Addressing these issues will require collaborative efforts from multiple disciplines beyond the HRI community.

3.1 The need for multidisciplinary definitions

Settling on a definition that specifies necessary and sufficient conditions of "overtrust" was challenging, but the workshop participants agreed that overtrust refers to situations where a robot or an automated system's expected performance exceeds its actual constraints. While the term conjured many different associations furthering concepts of trust, overtrust, and deception, reliance was ultimately considered paramount to such definition.

Trust, it should be noted, has long been understood as a relational construct [33]. Simmel's work on trust [34] offers an understanding built not only on the functional dimension of trust related to risk-taking and cooperation [35,36] but also on its embeddedness in systems of belief, particularly with reference to trust as a state of mind, or as faith. In the 1990s, scholars began to explore and

understand the connections between trust and social capital and focused on the development and maintenance of relationships and social order [37,38]. It will be critical to explore these dimensions of trust in relation to human–machine interaction and communication as humans develop new relationships with machines, particularly ones such as robots, digital voice assistants, and other kinds of intelligent agents [39–41].

Fundamentally, trust is a relational concept, based around vulnerability and the belief by an agent (*trustor*) that another entity (*trustee*) will do as expected. Both risk and interdependence are essential in sociological and economic conceptualizations of trust [42,43]. On one hand, risk is typically classified as the (perceived) probability of an uncontrolled loss. Without risk, trust is somewhat less relevant, as the trustor is not dependent on some other agent to reduce their own vulnerability to the said risk. With respect to risks, it has been suggested that a risk-based approach could be applied to prevent negative incidents [44]. On the other hand, interdependence is defined as a state in which one party has to rely on another to achieve a specific desired outcome. While these characteristics of trust pertain to trust in automation, trust in automation is characterized by additional factors that relate to the technology itself [45].

Some propose that the factors promoting trust in automation include the reliability, validity, and robustness of algorithmic systems [45,46]. These factors are said to strengthen the performance aspects of technology and promote trustworthy algorithmic systems. Trustworthy systems should be predictable (i.e. work the way they are supposed to), robust (withstand outside tampering), and able to fulfil the activity required of them. Since algorithmic systems are not single entities but interconnected elements that are deliberately or organically organized in a way that achieves a particular outcome, trust in automation includes not only factors related to performance but also those necessary for individuals to interact with them. These factors include traditionally human concepts, such as expertise and personality traits, environmental or cultural influences, the purpose(s) underlying an algorithmic system, as well as the intention(s) of those operating such systems [4,16,22,46]. These factors lead to trustable algorithmic systems or systems that a specified community accepts as competent, genuine, and/or honest.

3.2 Deception by robots

While potentially problematic for various reasons, we know that people trust robots in some contexts, and that robotic researchers actively cultivate trust-based relationships. Humans have evolved to depend on both information and practical support from other agents, which puts us at the risk of deception. If each individual had full and correct information in all situations, and the power to effect the changes they desired, they would not have to rely on others, and deception would not be as much of a problem as when we have to rely on veracity of the information provided by others and that they do the things they say – or we expect – them to do. In this context, it is useful to distinguish between epistemic trust in the quality of information we may acquire from robots, and pragmatic trust in a robot's physical capabilities. Trust in the quality of information partly requires that we trust the intentions and integrity of the sender of information, and while we will not assume that a robot is capable of having intentions, some people will in fact attribute such capabilities to robots. The discrepancy between what people believe a robot to be capable of (i.e. intentions and goodwill) and what a robot is actually capable of, has given rise to the notion that robots can be deceptive [47]. This will be examined in more detail below, as we examine different forms of robot deception. An example of such being robots that are designed to give the impression of empathy, with the goal of making a human believe it really has such capabilities, in order to elicit a certain response from users.

Some will argue that robots are not deceptive *per se*, as it is people's tendency to anthropomorphize technology that is problematic, i.e. our capacity to "imagine greater competence than is actually present" [48, p. 125]. In this view, people will deceive themselves into believing that just about anything is capable of having, for example, intentions and a purpose [49]. We argue that responsible social robotics requires designers to account for the human tendency to anthropomorphize technology. Robot deception may thus be intentional, or simply follow from negligence. As anthropomorphization of social robots can have effects on overtrust, it is necessary to not only be aware of active efforts to encourage this but also to proactively counteract unintended and unfortunate anthropomorphization [50].

Deception is central for the evolution of many intelligent systems. The most interesting question for HRI is often when and how a robot can be used for deception, rather than whether it *can* be deceptive *per se* [4,51]. Like in experimental psychology [52], deception on some level is arguably integral to research practices in social robotics. Quite often, researchers rely on a "manipulation technique" known as WoZ to control "the robot as a puppet to uncover specific social human behaviours when confronted with a machine" [53, p. 3]. Here, an ex-

perimenter or confederate operates the robot, without the research subject being aware, controlling various input variables in accordance with careful and meticulously described protocols for the particular study. In many cases, these kinds of simulations are necessary to learn about human–machine interactions, due to technological or other constraints [54]. There is nothing inherently suspicious or unacceptable with this research paradigm. But when an astute human agent is kept in the interactional loop, experimenters and robot-designers may also inadvertently create conditions for overattributing the robot's capabilities, which in turn may result in overtrust. In HRI this has necessitated rigorous reporting guidelines about WoZ protocols, and the role played by "wizards" in these interactions, including information about variables such as their demographics, training level, error rate, production- and recognition variables, and familiarity with the experimental hypothesis [54]. In terms of overtrust, these experimental practices may be consequential when social robots migrate from the cultural context of the research laboratory and into the outside world, like when social robots are considered implemented in healthcare practices. Just like with folk-concepts like "sociality," we need more empirical research on how trust and overtrust are created and maintained through joint, situated interactions between robots and their human interlocutors, as well as how design choices shape these encounters (see, for instance, [55]). From an ethical, legal, and social perspective, there is a need to understand the design choices that lead people to trust or distrust intelligent machines. Sometimes this means that we must extend our unit of analysis to encompass robot designers and experimenters in HRI (i.e. "wizards"). Insights about folk-theories of trust and overtrust that inform the design choices of roboticists and others involved in the making of robots can be acquired from ethnographic research of their work in naturalistic contexts (e.g. ref. [56]). One should, in other words, not heed the advice of The Wizard, if he booms: "pay no attention to that man behind the curtain!".[2]

Shim and Arkin provided a detailed examination of robot deception and the potential benefits of deception for HRI [57]. In doing so, they also provide a taxonomy of robot deception, with three dimensions (interaction object, goal, and type), each divided into two categories. As we later argue that deception impacts overtrust, their framework is useful for categorizing the target of deception, the reason for deception, and intentionality. For example, a robot can deceive a human or another robot (target), it can deceive in order to promote the perceived interests of the target or someone else (goal/reason), and it can deceive by physical or behavioural means (type) [57]. Another typology consists of three types: external, superficial, and hidden state deception [58]. The first relates to how a robot may deceive about other things than itself, such as a weather forecast. The second involves deception leading humans to believe the robot has capacities it does not, such as emotions, hopes, and desires. Hidden state deception occurs when we disguise actual features. These three categories are useful for categorizing the deception involved in encouraging or discouraging anthropomorphism, to which we shortly turn.

In closing, deception by robots is not inherently bad. More than that, it can be required for facilitating effective HRI, and it can even have positive consequences for the humans interacting with the robots [59,60]. While trust is important in human affairs, prosocial deception – for example white lies, games, and theatre – is also integral to our social interactions [61]. Robot deception can, in some situations, increase efficiency in care situations and levels of cooperation in games, increase human social engagement where the robot is present, and help build human–machine trust [50,62]. Furthermore, current social robots are not intentional agents, meaning that any deception is a byproduct of human design, whether intentionally or unintentionally. Not all deception is intentional, and robot deception can occur through coincidence or neglect [5]. Furthermore, deceptions are sometimes used by experimenters to elicit specific interactional effects when subjects are interacting with robots in the research laboratory. A designer that merely attempts to design a useful robot that uses, for example, referential and mutual gaze to improve the effectiveness of human interaction with the robot might end up creating a robot that deceives users into thinking it is something it is not. Our use of the concept of deception is thus based on a focus on the target of deception and not the intentions to deceive [5,63].

3.3 Does anthropomorphization lead to greater overtrust?

Social robots are increasingly embedded in numerous aspects of our lives, often at home or in the workspace. Their capabilities to mimic human speech, monitor our behaviours, and be part of social activities are social functions that all contribute to their "human likeness." However, these anthropomorphic features might obscure certain risks, for example concerning privacy [64], or

2 Line spoken by The Wizard of Oz, played by Frank Morgan, in the film The Wizard of Oz, directed by Victor Fleming (1939).

cloud our understanding that these robots are mere machines. While not robot-specific, anthropomorphism becomes highly relevant as robots increasingly mimic human features and capabilities [47,65].

This has led some to suggest that avoiding anthropomorphization is necessary for preventing overtrust [12]. These (perceived) blurred boundaries between humans and machines will make the overtrust phenomenon more relevant. Although people may be more vigilant in the early stages of introducing social robots, they may gradually be more willing to (over)trust a social robot that looks and sounds like them more easily than a smart home appliance, such as a vacuum [66,67].

Due to the complicated nature of the relationship between anthropomorphism and overtrust, the advice to avoid anthropomorphism appears to be too simplistic. Another possibility is to consider overtrust as a result of a lack of anthropomorphizing. When overtrust arises due to human goodwill, anthropomorphism may further exaggerate such tendencies. However, as mentioned, overtrust can be rooted in the robot's (real or idealized) capabilities, regardless of anthropomorphism. In these circumstances, using deception to actively encourage anthropomorphism might counter the tendency to overtrust. Anthropomorphism has essential effects on trust, as it can decrease initial expectations of robots and influence the human reaction to error [50]. As anthropomorphism entails projecting human characteristics onto robots, it might be used to encourage humans that interact with robots to perceive the robots as susceptible to error, much like humans. While anthropomorphism might help reduce overtrust caused by faith in the objectivity and infallibility of machines, alternatively, it could lead to other problems related to the perceived capabilities, for example, goodness or intent. In addition, we need to separate integrity-based overtrust from reliability-based. In the latter case, prosocial deception and the equivalent of robot white lies can enable designers to counteract overtrust, or at the very least, alleviate the adverse effects that stem from overtrust.

Solutions to avoid overtrusting robots have been sought. For example, deceptive practices could spur individuals to proactively engage in a situation rather than merely (over)relying on robots, thus encouraging the view of a robot as a social partner [68]. By equipping the robot with eyes, for example, the robot would appear to engage in mutual gaze and referential and deictic gaze in order to facilitate joint attention and encourage a set of beneficial human responses [62]. For instance, the social robot Pepper (Softbank) comes equipped with a random selection of micro-movements pertaining to eye gaze and hands, which create an impression that the machine is pragmatically competent. Mutual gaze increases trust and initiates joint attention, which could enable the robot to lead and divert human attention to important features of the environment, such as exit signs [69].

The extent to which humans will develop relationships with robots may also lead to potentially dangerous situations. The literature on trust and technology adoption indicates that users' expectations and experiences shape these encounters in fundamental ways, although some have hypothesized that human trust in robots may differ from other kinds of autonomous technologies [4]. Consequently, perceptions of robots, whether based on personal dispositions, direct experiences, or influence from widespread social norms and cultural representations in media, will likely form a complex baseline for trust in robots [70]. It is possible then to imagine scenarios in which users may, over a long period, interact with a robot in ways that cultivate trust, only to find that this trust breaks down at a particularly critical juncture, and should not be maintained in the future. Whether autonomous systems act as an empirical observer and reporter, or an interpreter of data and input (or possibly some combination of the two) will have consequences for trustworthiness.

3.4 Mapping the ecosystem

Robots and autonomous systems are far from being pure standalone objects that evolve within a clearly delineated space, but rather blur the traditional boundaries between products and services, disrupting existing paradigms [71]. As components in complex cyber–physical systems, robots, whether in the form of unmanned aircraft, autonomous transportation systems, or humanoid social robots, involve multiple actors and heavily interacting elements, which are highly connected and data reliant mediums [72]. As such, these socio-technological artefacts can only be apprehended and understood within a larger context, that is the social and technical ecosystem in which they evolve and through which they acquire their functions and values [73]. The notion of institution-based trust is sometimes used to describe this ecosystem and its relevance for interactions. In the context of online transactions and e-commerce, such institution-based trust is defined as "the belief that needed structural conditions are present (e.g. in the Internet) to enhance the probability of achieving a successful outcome in an endeavour like e-commerce" [74, p. 339] and acts as a precondition for more specific trusting beliefs (e.g. in an online vendor).

Building trust in associated ecosystems remains a central challenge to building trust in robotics generally [75]. Exact apprehensions of socio-technological settings in which they are deployed and used condition the way by which people leverage trust in them, as well as how their associated risks are assessed. While fostering trust in standalone objects themselves, such as a robot or some other autonomous system, remains critical for success of these robots or autonomous systems, it is equally important to extend trust to entire ecosystems. This latter point may be extremely challenging, given their great complexity and opaqueness from the perspective of most laypeople.

While fostering trust in robots is important for capitalizing on their value and societal benefits, mitigating overtrust in these complex artefacts will be just as crucial [76]. There may be potential misalignment between what human agents think a robotic system is capable of doing and its actual technological constraints. Overtrust in robots may also extend to the entire ecosystem in which they are embedded. For example, if users are sceptical towards institutionalized healthcare in a given country in general, they might also be sceptical towards healthcare robots. This could be problematic in the sense that users or the general public may never have an adequate understanding of a robot's functions and its place in the associated technological ecosystem, thereby resulting in a failure to grasp its inherent risks and potentially forgoing salient benefits.

A clear identification and mapping of different actors, components, and functions involved in robots and autonomous systems, as well as their relation to one another, is an essential prerequisite for assessing and addressing risks associated with the operation of social robots [72]. Great attention must therefore be given to delineating and understanding the layered interactions between socially situated robots and autonomous systems. Grasping the multileveled evolution of these robotic ecosystems, from the microlevel of human–machine interaction to the macrolevel of policy, legal, and economic systems, remains key for both fostering adequate trust and averting overtrust. Any policy or regulatory effort in this direction should be based on a holistic view of robots as contextually situated in a sea of social and technical values. Finally, from a liability perspective, mapping these ecosystems will greatly facilitate, at least in theory, the attribution of responsibility, should something go wrong with a given robot.

3.5 Overtrust and liability

The concept of trust also plays an important role with regards to liability for damage caused by robots [77–80]. Suppose, for instance, that an autonomous vehicle causes an accident because the "driver" did not pay attention to the road as they assumed the autonomous system would react when necessary (cf. accident with Tesla and Joshua Brown; see ref. [81]). Overtrusting the autonomous system of the vehicle led to the accident even though the system gave signals to the driver to take over the control again. The question that would arise is whether and under which circumstances the driver/user of the vehicle can be held liable for their fault under national law, for instance by violating a legal rule (e.g. traffic regulations) or acting negligently [80,82–85]. This will of course also depend upon the level of autonomy, for which the Society of Automotive Engineers' classification system, defining the degree of driving automation a car and its equipment may offer, is frequently used (SAE J3016).

Several challenges may arise under this fault-based liability, such as the question of whether the act of the vehicle can be attributed to the driver/user. Until now, it remains uncertain whether a user may actually trust the autonomous system in the vehicle. In Belgium, for instance, Article 8.3. of the Highway Code stipulates that the driver must at all times be able to perform the necessary driving actions and have his vehicle under control. The same level of attention is thus still required as from "traditional" drivers, which would undermine the benefits of autonomous traffic. The question also arises whether it is realistic to expect the same level of attention and the ability to actually induce the user/driver to intervene within (fraction of a) second. Autonomous vehicles will presumably be equipped with a system that warns users when they should take over the steering wheel. In case of an emergency, however, this has to be done in a couple of seconds or even a fraction thereof. Additionally, this requires the user's reactions to be adequate in these circumstances. Cognitive research shows that these assumptions are unrealistic [82,86,87].

Reliance on other legal regimes may thus be required. In this regard, objective liability regimes where the proof of fault is not required are viable options. Alternatively, reliance on the EU Product Liability Directive is possible [88,89]. A product is defective when it does not provide the safety which a person is entitled to expect (see chapters in ref. [79]). This is the so-called "legitimate expectations" test, where the concept of trust is at stake again. If a manufacturer emphasizes safety and trust into autonomous systems, there may be a higher risk of liability as the legitimate (safety) expectations of the general public increase [82,89]. The proper allocation of liability also impacts the question whether the user/driver can trust the vehicle, and thus also whether autonomous vehicles will become a reality/can be further commercialized. To fully realize

autonomous vehicles (and thus also the many benefits), the driver/user should actually be allowed to trust the vehicle without the risk of incurring liability. This means that supervision is no longer needed, and also implies that they should not/cannot be held liable when an error occurs, but instead another party will have to compensate for the damage, for instance, the insurance company, the operator, or the producer [80,84,85,90].

The concept of liability/allocation of liabilities will play an important role in creating trust towards vehicles and could increase autonomization of traffic. Liability is usually apprehended by actors as a risk that has to be mitigated. However, more broadly, liability can also be seen as a trust-building mechanism. For instance, if an efficient legal system of liability allocation exists, the victims obtain effective redress and the actors know when they could be exposed. Protection and legal certainty are, therefore, two elements for building trust in these new vehicles and robots globally. The specific issue of trust and autonomous vehicles is actually linked to the question of who should be held liable and it is important to think of liability as an instrument to regulate behaviour. After all, more research is needed on the question of whether liability really has a deterrent effect and incentivizes more careful conduct [91–93].

3.6 Education and literacy

Given that overtrust may become increasingly problematic with greater diffusion of autonomous systems, questions emerge about how users of such technologies can be adequately educated to prevent overtrust (which in turn can lead to overreliance, as an outcome of trust is often a "risk-taking behaviour" [94, p. 725], which includes use of and reliance on automation [2,16,18]). Should passengers of autonomous vehicles undergo dedicated safety training, in the sense of a license and test tailored to inform them about the capacities of the system at hand? What technical and operational knowledge is needed to retain a critical and rational stance towards the technology? Answers to these questions will be pressing in different institutional contexts such as mobility, healthcare, the law, households, and in schools/education.

For example, institutional review boards (IRBs) that oversee human participant research at universities and other organizations will need to learn about and be prepared for the use of AI, robots, and other digital interlocutors in research. While it bears noting that institutions are currently considering such advances, for example with respect to ethical concerns, these considerations are primarily aimed at physical safety or simple psychological harms, like the previously referenced deception studies. However, the unintended or longer term concerns such as trust/overtrust, and the consequences of the physical or psychological harms are ongoing fields of research, with many facets to explore. Similar to what occurred when scholars began to study the internet and online social phenomena it will take some time to explore and understand the consequences and nuances of both using robots *in* research and using robots *as* research instruments or collaborators [95]. In the former case, there are already many examples of research on HRI and human–machine communication that involve behavioural interventions between a human subject and a machine. An example might be a study undertaken to learn about subjects' preference for a robot's movement. In the latter case, a robot might be employed by a researcher to act as an interviewer or observer and collect data via audio or video recording and/or from other sensors. As was the case in the early days of internet research, it is likely that ethical guidelines will need to be developed and disseminated among researchers and IRBs to help guide understanding and evaluation of research protocols [96].

Finally, beyond user and oversight education, those determining critical decisions regarding liability in autonomous systems similarly should be educated in how systems operate and their impacts. For example, as concepts of liability/allocation of liabilities will play an important role in creating trust, both legislators and the judiciary will need to obtain a level of technological understanding to appropriately regulate and enforce measures maintaining safety and preserving community values. Without such understanding there may be over-regulation, creating a chilling effect on innovation, or under-regulation where individuals no longer trust technological advancements. While regulation will always be one step behind innovation, understanding current systems and where developments are heading can help keep critical players informed and ready to tailor bespoke policies. Though scholars and government entities (e.g. the EU's Expert Group on Liability and New Technologies) alike have begun to address major gaps and issues within liability and regulatory schemes applicable to autonomous systems and robotics, ongoing research is needed to modernize existing regulatory regimes, as well as to educate regulators in order to update and propose appropriate safeguards [97–100].

4 Conclusion

In this contribution, we offered an overview of pressing topics in the context of overtrust and robots. In our review, we mobilized insights solicited from in-depth

conversations from a multidisciplinary workshop on the topic, held at a leading robotics conference. The participants brought in their expertise, which allowed us to formulate a forward-looking research agenda.

In light of the increasing use of automated systems, both embodied and disembodied, overtrust is becoming an ever more important topic. However, our overview shows how the overtrust literature has so far been mostly confined to HRI research and psychological approaches. While philosophers, ethicists, engineers, lawyers, and social scientists more generally have a lot to say about trust and technology, conceptual clarity and a shared framework for approaching overtrust are missing. In this article, our goal was not to provide an overarching framework but rather to encourage further dialogue from an interdisciplinary perspective, integrating diverse literature and providing a ground for common understanding. Our research agenda centres around the following six topics:

First, the need for multidisciplinary definitions: While we loosely converged on a mutual understanding of overtrust as a gap between the real constraints of an automated system and its expected performance, we had to acknowledge that trust is a very complex and elusive construct. Scholars have offered dozens of definitions of trust, stressing aspects such as risk, vulnerability, (lack of) control, and expectation. The plurality of (sometimes diverging) understandings naturally complicates a common ground. We agreed that a relational perspective that focuses quite strongly and descriptively, but not exclusively, on rational dynamics and reliance is fruitful. The need for multidisciplinary definitions will stay relevant in the future and empirical research could try to systematically map and consolidate existing scholarship (e.g. through a systematic literature review on overtrust).

Second, addressing deception by robots: Beyond conceptual clarification, understanding the link between overtrust and deception is a key aspiration for the community. Deception has emerged as an important topic in robotics research. However, it is not necessarily bad, as we have argued that deception may both be a cause of and cure for overtrust in robots. However, what are the short-term and long-term implications if deception deliberately exploits people's overtrust in robots? Robot deception may lead to a corrosion of trust and cooperation between humans [5], an argument in line with Danaher's examination of robots as catalysts for moral revolutions [58]. Regardless of the long-term consequences, the short-term consequences of robot deception are both real and important, and more attention should be directed towards understanding and mitigating the potential negative consequences of such deception [63]. Simultaneously, deception may both be unavoidable and beneficial as well, which suggests a need to focus on balancing the good against the good in this case.

Third, addressing anthropomorphization: In social robotics, anthropomorphization is an important research theme, showing the close relation between a robot's design, embodiment, and its affordances, including its perception with aspects such as social bonding, social presence, trust, and overtrust. We have discussed the ambivalence of anthropomorphization and overtrust. Users might trust human-like robots too much because they are perceived as sentient, and such sentience might be construed as reliable, benevolent, and honest. By contrast, users might trust machine-like robots too much because they see machines as more capable and unbiased than humans. With complex systems, both a tendency for overreliance in the form of a machine heuristic [101] and underreliance in the form of algorithm aversion [102] have been demonstrated. Future research ought to assess under which conditions anthropomorphization produces overtrust.

Fourth, mapping the ecosystem: Robots and autonomous systems are at the crossroads of products and services and embody a complex landscape of actors and cyber–physical components which interact with each other [71,72]. Fostering trust in such smart and connected artefacts/systems should not be limited to the standalone object or service itself, but should rather cover the entire socio-technical ecosystem in which they are embedded. Conversely, the issue of overtrusting robots and autonomous systems should also extend to this complex socio-technical environment. It is therefore important for further research to better identify and map the numerous actors, components, and functions involved in robots and autonomous systems, as well as more clearly delineate and understand the multilayered interactions within these ecosystems. Such an endeavour will be crucial for assessing the inherent risks associated with these robots and autonomous systems. This should allow to both adequately foster societal trust and help bridge the gap between the perceived and actual risks, therefore mitigating potential overtrust invested in them.

Fifth, liability: From a legal perspective, overtrust in technology is closely tied to questions of liability – or who should be responsible for addressing the damage caused by a robot or autonomous system. Our analysis has shown that existing liability regimes are partly suitable to deal with overtrust in robots, including autonomous vehicles. Fault-based liability regimes, however, might not be sufficient. Alternative liability regimes, such as the EU Product Liability Directive or liability for defective items will gain importance [89,90]. Future research could look more closely at the interplay between

(over)trust and legal responsibility. An evaluation of current liability mechanisms' failures/limitations in leveraging an optimal level of trust in these disruptive artefacts should be considered. Conversely, an assessment of the liability implications of overtrusting such technologies should be carried out. Finally, in light of these assessments, and pending findings, further research could extend to considering normative evolutions which would address some of the identified hurdles.

Sixth, education and literacy: Finally, we identified a need to include the topic of overtrust more strongly in curricula and literacy initiatives. Robotics programs at higher education institutions could integrate some of the aspects discussed above in their curricula to reflect consciously on overtrust in design. However, literacy should also be promoted in other settings where robots and autonomous systems are developed, employed, and regulated. User manuals could stress overtrust as a safety risk and convey accessible information on the benefits as well as downsides of overtrust with the specific robot. From a research ethics point-of-view, new considerations arise when robots are used for data collection as relatively autonomous agents (in the sense of an interviewer or social scientist). IRBs should familiarize themselves with the latest developments in this regard. Finally, regulators and policymakers should likewise be prepared for the level of sophistication required to appropriately analyse such emerging technologies.

Our interest in overtrust aligns with the European Commission's focus on trust and trustworthiness in the context of AI [1]. Technical, legal, and social solutions to overtrust should be given joint consideration as technologies such as autonomous vehicles enter the mainstream. From a technological standpoint, value-sensitive design could prevent users from overtrusting robots and show the inherent uncertainties and constraints in autonomous decision-making. In addition, teachings from research on transparency and trust in the context of robots and AI could come in handy [8,32,103]. From a legal/policy perspective, diligent certification, updating liability regimes with clear mechanisms, and consistently applying sectoral regulation are crucial to ensuring that overtrust does not become too much of a problem. While current legal regimes are ultimately capable of handling harms caused by robots, modernization in interpretation and regulation are crucial to continue evolving and responding to the legal challenges posed by robotics. Finally, and from a social point-of-view, literacy initiatives, as discussed in previous sections, and meaningful expectation management among robot vendors and operators could mitigate overtrust. All in all, addressing overtrust in HRI demands a multi-level analysis ranging from the technical, individual, and societal levels. Consequently, further research efforts examining overtrust should comprehensively address the interplay between different levels as such an undertaking is essential to understand the impact of this phenomenon on society holistically.

Funding information: Orian Dheu received funding from the European Union's Horizon 2020 Research and Innovation – Marie Skłodowska-Curie actions program under the Safer Autonomous Systems (SAS) project, grant agreement no. 812.788: https://etn-sas.eu/. This publication reflects only the author's view, exempting the European Union from any liability. Eduard Fosch-Villaronga received funding from the LEaDing Fellows Marie Curie COFUND fellowship, a project funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 707404. Aurelia Tamò-Larrieux received funding from the International Postdoctoral Fellowship grant, University of St. Gallen, Switzerland: project number 1031564. Christoph Lutz received funding from the Research Council of Norway under grant agreement 275347 "Future Ways of Working in the Digital Economy." Mads Solberg's contribution was supported by grant no. 285216 from the Regional Research Fund of Central-Norway. Researchers of the SIRRL Laboratory at University of Waterloo are supported in part thanks to funding from the Canada 150 Research Chairs Program.

Author contributions: The authors are listed in alphabetical order.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

[1] European Commission, "On artificial intelligence – a European approach to excellence and trust," *European Commission*. Available https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[2] J. D. Lee and K. A. See, "Trust in automation: designing for appropriate reliance," *Hum. Factors J. Hum. Factors Ergonom Soc.*, vol. 46, no. 1, pp. 50–80, 2004.

[3] D. Meyerson, K. E. Weick, and R. M. Kramer, "Swift trust and temporary groups," in *Trust Organizations: Frontiers of Theory and Research*, T. Tyler, Ed., California, US: SAGE Publications, Inc, 1996, pp. 166–195.

[4] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of

factors affecting trust in human-robot interaction," *Hum. Factors*, vol. 53, no. 5, pp. 517–527, 2011.
[5] H. S. Sætra, "Social robot deception and the culture of trust," *Paladyn J. Behav. Robot.*, vol. 12, no. 1, pp. 276–286, 2021.
[6] E. E. Levine and M. E. Schweitzer, "Prosocial lies: When deception breeds trust," *Organ. Behav. Hum. Decis. Process.*, vol. 126, pp. 88–106, 2015.
[7] S. C. Robinson, "Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI)," *Technol. Soc.*, vol. 63, art. 101421, 2020.
[8] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns," *Big Data Soc.*, vol. 6, no. 1, pp. 1–14, 2019.
[9] R. Parasuraman and V. Riley, "Humans and automation: use, misuse, disuse, abuse," *Hum. Factors*, vol. 39, no. 2, pp. 230–253, 1997.
[10] High Level Expert Group on AI, "Ethics guidelines for trustworthy AI," European Commission, 2020. Available: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1
[11] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.
[12] A. R. Wagner, J. Borenstein, and A. Howard, "Overtrust in the robotic age," *Commun. ACM*, vol. 61, no. 9, pp. 22–24, 2018.
[13] K. L. Mosier, M. Dunbar, L. McDonnell, L. J. Skitka, M. Burdick, and B. Rosenblatt, "Automation bias and errors: Are teams better than individuals?," *Proc. Hum. Factors Ergonom. Soc. Ann. Meet.*, vol. 42, no. 3, pp. 201–205, 1998.
[14] M. Itoh, "Toward overtrust-free advanced driver assistance systems," *Cognit. Technol. Work.*, vol. 14, no. 1, pp. 51–60, 2012.
[15] M. Moray and T. Inagaki, "Attention and complacency," *Theor. Issues Ergonom. Sci.*, vol. 1, no. 4, pp. 354–365, 2000.
[16] K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation," *Hum. Factors*, vol. 58, no. 3, pp. 377–400, 2016.
[17] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Machine Stud.*, vol. 27, no. 5–6, pp. 527–539, 1987.
[18] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Human-Computer Stud.*, vol. 58, no. 6, pp. 697–718, 2003.
[19] M. Itoh, G. Abe, and K. Tanaka, "Trust in and use of automation: their dependence on occurrence patterns of malfunctions," in *IEEE SMC'99 Conference Proceedings, 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, vol. 3, 1999, pp. 715–720.
[20] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.
[21] I. Gaudiello, E. Zibetti, S. Lefort, M. Chetouani, and S. Ivaldi, "Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers," *Comput Hum. Behav.*, vol. 61, pp. 633–655, 2016.
[22] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015.
[23] J. J. Trinckes, Jr., *The Definitive Guide to Complying with the HIPAA/HITECH Privacy and Security Rules*, Boca Raton, FL: CRC Press, 2013.
[24] K. D. Mitnick, *The Art of Deception: Controlling the Human Element of Security*, New York: Wiley, 2003.
[25] B. Postnikoff and I. Goldberg, "Robot social engineering," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018.
[26] S. Booth, J. Tompkin, H. Pfister, J. Waldo, K. Gajos, and R. Nagpal, "Piggybacking robots," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017.
[27] A. M. Aroyo, F. Rea, G. Sandini, and A. Sciutti, "Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3701–3708, 2018.
[28] A. M. Aroyo, T. Kyohei, T. Koyama, H. Takahashi, F. Rea, A. Sciutti, et al., "Will people morally crack under the authority of a famous wicked robot?," in *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018.
[29] J. Borenstein, A. R. Wagner, and A. Howard, "Overtrust of pediatric health-care robots: A preliminary survey of parent perspectives," *IEEE Robot. Autom. Mag.*, vol. 25, no. 1, pp. 46–54, 2018.
[30] European Commission, Proposal for a Regulation on AI, AI Act, 2021. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
[31] A. Martinetti, P. Chemweno, K. Nizamis, and E. Fosch-Villaronga, "Redefining safety in light of human-robot interaction: A critical review of current standards and regulations," *Front. Chem. Eng.*, vol. 3, art. 666237, 2021.
[32] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Towards transparency by design for artificial intelligence," *Sci. Eng. Ethics*, vol. 26, no. 6, pp. 3333–3361, 2020.
[33] G. Becker, "A theory of social interactions," *J. Political Econom.*, vol. 82, pp. 1063–1091, 1974.
[34] G. Simmel, *The Sociology of Georg Simmel*, New York: Free Press, 1950.
[35] J. S. Coleman, *Foundations of Social Theory*, Cambridge, MA: Harvard University Press, 1990.
[36] D. Gambetta, *Trust: Making and Breaking Co-operative Relations*, Oxford: Basil Blackwell, 1988.
[37] P. Bourdieu, "The forms of capital," in *Education: Culture, Economy, and Society*, A. H. Halsey, H. Lauder, P. Brown, and A. S. Wells, Eds, Oxford: Oxford University Press, 1997.
[38] F. Fukuyama, *Trust: The Social Virtues and the Creation of Prosperity*, London: Hamish Hamilton, 1995.
[39] S. Jones, "People, things, memory and human-machine communication," *Int. J. Media Cult. Politics*, vol. 10, no. 3, pp. 245–258, 2014.
[40] A. L. Guzman, "The messages of mute machines: Human-machine communication with industrial technologies," *Commun. + 1*, vol. 5, no. 1, art. 4, 2016.

[41] A. L. Guzman, "What is human-machine communication, anyway," in *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves*, A. Guzman, Ed., Bern: Peter Lang, 2018, pp. 1–28.
[42] E. Keymolen, "Opinions when cities become smart, is there still place for trust?," *Eur. Data Prot. Law Rev.*, vol. 5, no. 2, pp. 156–159, 2019.
[43] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Acad. Manag. Rev.*, vol. 23, no. 3, pp. 393–404, 1998.
[44] G. M. Bounds and N. Malyshev, *OECD Reviews of Regulatory Reform Risk and Regulatory Policy: Improving the Governance of Risk*, Paris: OECD Publishing, 2010.
[45] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink, "Trust in automation," *IEEE Intell. Syst.*, vol. 28, no. 1, pp. 84–88, 2013.
[46] K. Siau and W. Wang, "Building trust in artificial intelligence, machine learning, and robotics," *Cut. Bus. Technol. J.*, vol. 31, no. 2, pp. 47–53, 2018.
[47] A. Sharkey and N. Sharkey, "Children, the elderly, and interactive robots," *IEEE Robot. Autom. Mag.*, vol. 18, no. 1, pp. 32–38, 2011.
[48] S. C. Levinson, "Natural forms of purposeful interaction among humans: What makes interaction effective?," in *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions*, K. A. Gluck, J. E. Laird, Eds, Cambridge, MA: MIT Press, 2019, pp. 111–126.
[49] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media like Real People*, Cambridge, United Kingdom: Cambridge University Press, 1996.
[50] E. J. de Visser, S. S. Monfort, R. McKendrick, M. A. Smith, P. E. McKnight, F. Krueger, et al., "Almost human: Anthropomorphism increases trust resilience in cognitive agents," *J. Exp. Psychol Appl.*, vol. 22, no. 3, pp. 331–349, 2016.
[51] A. R. Wagner and R. C. Arkin, "Acting deceptively: Providing robots with the capacity for deception," *Int. J. Soc. Robot.*, vol. 3, no. 1, pp. 5–26, 2011.
[52] R. Hertwig and A. Ortmann, "Deception in experiments: Revisiting the arguments in its defense," *Ethics Behav.*, vol. 18, no. 1, pp. 59–92, 2008.
[53] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan and T. Belpaeme, "From characterising three years of HRI to methodology and reporting recommendations," in *11th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), 2016.
[54] L. Riek, "Wizard of Oz studies in HRI: a systematic review and new reporting guidelines," *J. Human-Robot Interact.*, vol. 1, no. 1, pp. 119–136, 2012.
[55] M. Alač, J. Movellan, and F. Tanaka, "When a robot is social: Spatial arrangements and multimodal semiotic engagement in the practice of social robotics," *Soc. Stud. Sci.*, vol. 41, no. 6, pp. 893–926, 2011.
[56] M. Alač, "Moving android: On social robots and body-in-interaction," *Soc. Stud. Sci.*, vol. 39, no. 4, pp. 491–528, 2009.
[57] J. Shim and R. C. Arkin, "A taxonomy of robot deception and its benefits in HRI," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013.
[58] J. Danaher, "Robot betrayal: A guide to the ethics of robotic deception," *Ethics Inf. Technol.*, vol. 22, no. 2, pp. 117–128, 2020.
[59] M. Coeckelbergh, "Are emotional robots deceptive?," *IEEE Trans. Affect. Comput.*, vol. 3, no. 4, pp. 388–393, 2011.
[60] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! An interaction with a cheating robot," in *5th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), 2010.
[61] E. E. Levine and M. E. Schweitzer, "Prosocial lies: When deception breeds trust," *Organ. Behav. Hum. Decis. Process.*, vol. 126, pp. 88–106, 2015.
[62] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *J. Human-Robot Interact.*, vol. 6, no. 1, pp. 25–63, 2017.
[63] A. Sharkey and N. Sharkey, "We need to talk about deception in social robotics!," *Ethics Inf. Technol.*, pp. 1–8, 2020. doi: 10.1007/s10676-020-09573-9.
[64] C. Lutz, M. Schöttler, and C. P. Hoffmann, "The privacy implications of social robots: Scoping review and expert interviews," *Mob. Media Commun.*, vol. 7, no. 3, pp. 412–434, 2019.
[65] K. Darling, "'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy," *Robot Ethics 2.0*, P. Lin, K. Abney, and R. Jenkins, Eds, Oxford: Oxford University Press, 2017, pp. 173–192.
[66] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The physical presence of a robot tutor increases cognitive learning gains," *Proc. Annu. Meet. Cognit. Sci. Soc.*, vol. 34, pp. 1882–1887, 2012.
[67] C. Sinoo, S. van der Pal, O. A. Blanson Henkemans, A. Keizer, B. P. B. Bierman, R. Looije, et al., "Friendship with a robot: Children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management," *Patient Educ. Counseling*, vol. 101, no. 7, pp. 1248–1255, 2018.
[68] H. S. Sætra, "The parasitic nature of social AI: Sharing minds with the mindless," *Integr. Psychol Behav. Sci.*, vol. 54, no. 2, pp. 308–326, 2020.
[69] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *J. Human-Robot Interact.*, vol. 6, no. 1, pp. 25–63, 2017.
[70] K. S. Haring, C. Mougenot, F. Ono, and K. Watanabe, "Cultural differences in perception and attitude towards robots," *Int. J. Affect. Eng.*, vol. 13, no. 3, pp. 149–157, 2014.
[71] E. Fosch-Villaronga and C. Millard, "Cloud robotics law and regulation," *Robot. Autonom Syst.*, vol. 119, pp. 77–91, 2019.
[72] C. Lutz and A. Tamò, "Communicating with robots: ANTalyzing the interaction between healthcare robots and humans with regards to privacy," in *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves*, A. Guzman, Ed., Bern: Peter Lang, 2018, pp. 145–165.
[73] A. Hepp, "Artificial companions, social bots and work bots: communicative robots as research objects of media and communication studies," *Media Cult. Soc.*, vol. 42, no. 7–8, pp. 1410–1426, 2020.
[74] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Inf. Syst. Res.*, vol. 13, no. 3, pp. 334–359, 2002.
[75] B. C. Kok and H. Soh, "Trust in robots: challenges and opportunities," *Curr. Robot. Rep.*, vol. 1, no. 4, pp. 297–309, 2020.

[76] A. Howard and J. Borenstein, "Trust and bias in robots," *Am. Scientist*, vol. 107, no. 2, p. 86, Mar–Apr 2019.
[77] W. Barfield, "Liability for autonomous and artificially intelligent robots," *Paladyn, J. Behav. Robot.*, vol. 9, no. 1, pp. 193–203, 2018.
[78] M. Ebers and S. Navas, *Algorithms and Law*, Cambridge: Cambridge University Press, 2020.
[79] S. Lohsse, R. Schulze, and D. Staudenmayer, *Liability for Artificial Intelligence and the Internet of Things*, Baden-Baden: Nomos, 2019.
[80] E. Tjong Tjin Tai, "Liability for (semi)autonomous systems: robots and algorithms," in *Research Handbook on Data Science and Law*, V. Mak, E. Tjong Tjin Tai, and A. Berlee, Eds, Cheltenham: Edward Elgar, 2018, pp. 55–82.
[81] R. Abrams and A. Kurtz, "Joshua Brown, who died in self-driving accident, tested limits of his tesla," *The New York Times*, July 1, 2016. Available https://www.nytimes.com/2016/07/02/business/joshua-brown-technology-enthusiast-tested-the-limits-of-his-tesla.html.
[82] J. De Bruyne and J. Tanghe, "Liability for damage caused by autonomous vehicles: A Belgian perspective," *J. Eur. Tort Law*, vol. 8, no. 3, pp. 324–371, 2018.
[83] M. Schellekens, "Self-driving cars and the chilling effect of liability law," *Comp. Law Secur. Rev.*, vol. 31, no. 4, pp. 506–517, 2015.
[84] H. Surden and M. A. Williams, "Technological opacity, predictability, and self-driving cars," *Cardozo Law Rev.*, vol. 38, pp. 121–181, 2016.
[85] N. Velinga, "Legal aspects of automated driving: on drivers, producers, and public authorities," *PhD thesis*, University of Groningen, Groningen, Netherlands, 2020.
[86] T. Malengreau, "Automatisation de la conduite: quelles responsabilités en droit belge?," *RGAR*, vol. 5, pp. 15578–15607, 2019.
[87] J. De Bruyne, *Autonome Motorvoertuigen*, Bruges: Vanden Broele, 2021.
[88] K. Funkhouser, "Paving the road ahead: autonomous vehicles, products liability, and the need for a new approach," *Utah Law Rev.*, vol. 1, pp. 437–462, 2013.
[89] J. Werbrouck, "De productaansprakelijkheid voor zelfrijdende motorrijtuigen," *TPR*, vol. 55, pp. 529–604, 2018.
[90] D. Levalley, "Autonomous vehicle liability – application of common carrier liability," *Seattle Univ. Law Rev.*, vol. 36, pp. 5–26, 2013.
[91] G. Calabresi, *The Costs of Accidents: A Legal and Economic Analysis*, New Haven: Yale University Press, 1970.
[92] J. De Bruyne, *Third-Party Certifiers*, Alphen aan den Rijn: Kluwer Law International, 2019.
[93] R. A. Posner, "A Theory of Negligence," *J. Leg. Stud.*, vol. 1, no. 1, pp. 29–96, 1972.
[94] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Acad. Manag. Rev.*, vol. 20, no. 3, pp. 709–734, 1995.
[95] S. Jones, "Can social robots do social science? The ethics of robots in research," *NordiCHI Conference*, Oslo, 2018.
[96] a. s. franzke, A. Bechmann, M. Zimmer, C. Ess, and the Association of Internet Researchers, "Internet Research: Ethical Guidelines 3.0," *Association of Internet Researchers* (*AoIR*), 2020. Available https://aoir.org/reports/ethics3.pdf
[97] A. Bertolini, "Robots as products: The case for a realistic analysis of robotic applications and liability rules," *Law, Innov. Technol.*, vol. 5, no. 2, pp. 214–247, 2013.
[98] R. Calo, "Robotics and the lessons of cyberlaw," *Calif. Law Rev.*, vol. 103, no. 3, pp. 513–563, 2015.
[99] C. Gordon and T. Lutz, "Haftung für automatisierte Entscheidungen – Herausforderungen in der Praxis," *Schweizerische Z. für Wirtschafts- und Finanzmarktrecht*, vol. l, pp. 53–61, 2020.
[100] M. F. Lohmann, "Roboter als Wundertüten – eine zivilrechtliche Haftungsanalyse," *Aktuelle juristische Praxis: AJP*, vol. 2, pp. 152–162, 2017.
[101] S. S. Sundar and J. Kim, "Machine heuristic: When we trust computers more than humans with our personal information," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.
[102] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *J. Exp. Psychology: Gen.*, vol. 144, no. 1, pp. 114–126, 2014.
[103] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Robots and transparency: The multiple dimensions of transparency in the context of robot technologies," *IEEE Robot. Autom. Mag.*, vol. 26, no. 2, pp. 71–78, 2019.

View publication stats