Regulation &
Governance

*Regulation & Governance* (2023) doi:10.1111/rego.12568

# Regulating for trust: Can law establish trust in artificial intelligence?

Aurelia Tamò-Larrieux
*Law & Tech Lab, Maastricht University, Maastricht, The Netherlands*

Clement Guitton, Simon Mayer
*Institute of Computer Science, University of St. Gallen, St. Gallen, Switzerland*

Christoph Lutz
*BI Norwegian Business School, Oslo, Norway*

**Abstract**

The current political and regulatory discourse frequently references the term “trustworthy artificial intelligence (AI).” In Europe, the attempts to ensure trustworthy AI started already with the High-Level Expert Group Ethics Guidelines for Trustworthy AI and have now merged into the regulatory discourse on the EU AI Act. Around the globe, policymakers are actively pursuing initiatives—as the US Executive Order on Safe, Secure, and Trustworthy AI, or the Bletchley Declaration on AI showcase—based on the premise that the right regulatory strategy can shape trust in AI. To analyze the validity of this premise, we propose to consider the broader literature on trust in automation. On this basis, we constructed a framework to analyze 16 factors that impact trust in AI and automation more broadly. We analyze the interplay between these factors and disentangle them to determine the impact regulation can have on each. The article thus provides policymakers and legal scholars with a foundation to gauge different regulatory strategies, notably by differentiating between those strategies where regulation is more likely to also influence trust on AI (e.g., regulating the types of tasks that AI may fulfill) and those where its influence on trust is more limited (e.g., measures that increase awareness of complacency and automation biases). Our analysis underscores the critical role of nuanced regulation in shaping the human-automation relationship and offers a targeted approach to policymakers to debate how to streamline regulatory efforts for future AI governance.

**Keywords:** artificial intelligence, automation, human-automation trust, regulation of technology, trust, trustworthy AI.

## 1. Introduction

In May 2021, the EU Executive Vice-President, Margrethe Vestager, held a speech in which she asked: “Do Europeans trust technology?” After noting several large discrepancies within member states but an overall low level of trust, she offered four solutions to increase trust: via *educating* users, and via three *regulations*, namely on the single market, public services, and digital rights (Vestager, 2021). In effect, Vestager is suggesting that such regulations can modulate trust levels.[1] She is not alone in doing so and many other EU documents share this rationale, implicitly or explicitly: For instance, in October 2021, the European Commission issued a call to adapt liability regulations to AI which stated that “[t]he Commission’s objective is to encourage the development and roll-out of safe AI systems and build trust among potential users.”[2]

The drivers behind the focus on trust in automation,[3] and more specifically in AI, are plentiful: The rapid development of AI-based tools and embodiment of technology (e.g., within social robots) has changed our relationship with technology. Today, technological artifacts are more and more perceived as “interaction partners” or “interlocutors” rather than services and we interact with technology in a more intimate and social manner (Guzman & Lewis, 2020; NIST, 2021). This shift in perception and interaction has also led to a rich discourse in

Correspondence: Aurelia Tamò-Larrieux, Law & Tech Lab, Maastricht University, Bouillonstraat 1-3, 6211 LH Maastricht, The Netherlands. Email: a.tamo@maastrichtuniversity.nl

Accepted for publication 14 November 2023.

academia on trust in automation and AI (Glikson & Woolley, 2020; Kaur et al., 2022; Yang & Wibowo, 2022) and within the political discourse (HLEG, 2019; see Section 4). Particularly, policymakers are fearful that without proper regulation, society will trust AI too little to reap its benefits or, on the contrary, that society might trust AI too much (especially in case of errors or AI hallucinations). In short, these concerns center around the notion of *distrust* and, at the other end of the spectrum, *overtrust*. While distrust does not refer to a lack of trust but to a confident expectation that the interaction will have a negative consequence (Hill & O'Hara, 2006), overtrust describes a situation where a person misunderstands or miscalibrates the uncertainty and vulnerability of a certain interaction and misplaces trust into a trusted agent (Aroyo et al., 2021). Considering this, the focus on regulating "for" trust is not surprising, but the baseline assumption that regulation can directly establish trust must be challenged. To address the question of the extent to which regulation can foster trust in AI, we take stock of research on factors influencing trust—including by noting differences in trust in automation.

Trust in general, and trust in automation (including AI) in particular, has been subjected to much research, leading to different conceptualizations and empirical models. Many different definitions of trust have emerged in the literature (Bodó, 2021; Botsman, 2017; Hardin, 2002; Hult, 2018; Keymolen, 2016; Lee & See, 2004; Möllering, 2006; Rousseau et al., 1998; Sztompka, 1999), typically centering around core features of trust. First, trust involves at least two agents and typically an action that one agent has to fulfill. Second, the relationship and situation in which this action has to occur is characterized by uncertainty and vulnerability by the agent that should benefit from the action. Third, there is confidence of the trusting agent that the other agent will act according to the goals of the trusting agent. Thus, trust boils down to this expectation or attitude (i.e., a cognitive concept, as described by Baier, 1986; Hardin, 2002) of the trusting agent that the trusted agent will act in a predetermined manner. The literature on trust in general, and trust in automation in particular, highlights how complex the topic is, with core conditions and various factors influencing the human-automation trust relationship. We conceptualize this complexity by outlining 16 propositions that emerge from the literature and describe aspects that influence trust. Within these propositions, we elaborate on research findings in the field of trust and automation. Building upon this framework, we address the question of how law, or regulation more broadly, impacts the described factors in a way that enables trust. The answer to this question has clear policy implications, as striving for trust can have on one extreme a positive connotation (aiming for a high, aspiring objective) and on the other extreme a negative one (focusing on a wrongheaded approach that is bound to fail).

The article is structured as follows: In Section 2, we discuss how trust and regulation interplay. We analyze the extent to which regulation can influence trust by discussing literature on how law capitalizes on trust and minimizes distrust. We show that reducing every known complexity of an interaction is unrealistic, contradicting arguments that regulation can completely replace trust by eliminating uncertainty. Thus, the role of regulation becomes one of creating a framework that renders the uncertainty and vulnerability of interactions manageable for society. Upon this basis, we provide a theoretical model for trust in Section 3, which captures the complexity of trust in automation by describing the core factors impacting trust. We elaborate on the impact of these factors on trust in 16 propositions, building upon the rich literature in the field of human-automation and human-human trust. We also discuss if and how these factors can be shaped by regulation, analyzing the degree of impact of the propositions on human-automation trust relationships. We apply our model to the recently adopted AI Act of the EU[4] in Section 4 and highlight how different articles within the AI map to the propositions elaborated in Section 3. In Section 5, we discuss our main findings and conclude with some key implications.

## 2. On the interplay of trust and regulation

When analyzing the interplay between law and trust, a key question is: To what extent can regulation influence trust? Interestingly, there is a dearth of (legal) research analyzing the interplay between law and trust, or how law capitalizes on trust and minimizes distrust (see, e.g., Blair & Stout, 2000; Cross, 2004; Hill & O'Hara, 2006; Hult, 2018). Still, the existing literature can provide insights into related questions. For instance, Hill and O'Hara (2006, p. 1718) ask: "To what extent do legal rules, cases, and law enforcement efforts enhance or detract from the trust present in relationships?" Two sub-questions arise: One on *co-existence*, the other on *optimization*.

### 2.1. Law and trust: Co-existence

First, the extent of the impact of law on trust must be understood and measured. The central question here revolves around: To what extent can trust and law co-exist? Legal literature has addressed this question from different perspectives. Some have focused on the reduction of uncertainty created by law (Hill & O'Hara, 2006). Others have analyzed whether law erodes the need for trust, concluding that the law can minimize the negative consequences of overtrust (Cross, 2004). And yet another group of researchers have contrasted the benefits of trust for cooperation with the downsides of establishing complex legal agreements (Blair & Stout, 2000).

The premise of the trust paradox is that if one establishes assurance structures in order to enable trust, the need for trust is consequently diminished (Cheshire, 2011). The more assurances are built up, the more uncertainty is reduced, and the less there is a need for trust. In the literature, it is however unclear to what extent a mere reduction of uncertainty will lead to a trust paradox, most notably: Is full certainty of the outcomes necessary, or is the ability to have redress mechanisms to ensure a specific outcome sufficient (notwithstanding the cost of making use/enforcing these mechanisms)? For instance, Cheshire (2011) argues that when strict legal rules are established, including robust monitoring systems, cooperation among parties is present but this cooperation is not based on trust but rather on the fear of sanctions.

The trust paradox has also been analyzed in organizational studies on the effect of control on trust (Long & Sitkin, 2018). A special focus has rested on *contracts* among parties as a control measure (Deakin et al., 1994). Here the literature has generated mixed results with respect to the question of whether contracts act as complements or substitutes to trust, showing that the use of contracts can act as a substitute for trust, but can also have a negative impact on the trust relationship (Long & Sitkin, 2018). Long and Sitkin (2018) show that an adequate balance between the use and enforcement of contracts and no formal agreements can be struck to generate productive trusting relationships among co-workers and managers (Long & Sitkin, 2018). However, an important gap in this literature is still the one-sided focus on either the form of control (e.g., contracts, surveillance) or on the ways/application of control (the "how"; e.g., strictly, fairly, etc.), yet both the "what" and the "how" matter simultaneously (Long & Sitkin, 2018).

In the legal literature, we find different stances with respect to the co-existence of trust and law. According to Ribstein (2001), law substitutes trust. He argues that law is neither helping nor supporting the development of trust but that on the contrary, "the shadow of coercion" impedes such trust from developing altogether (Ribstein, 2001, p. 564). Hult (2018) offers a more nuanced approach to the crowding out effect of regulation. Here, coercion in the form of sanctions pushes a party to (re)assess the probability attached to negative outcomes and, based on this assessment, interact with another party. Ribstein's (2001) claim, however, seems short-sighted, as it assumes that through the law, the uncertainty is reduced to such an amount that the probabilities of harm occurring can not only be calculated but also rationally be taken into consideration. The former is unlikely, especially in complex interactions among individuals, institutions, and automation systems, while the latter assumes the prevalence of rational agents, which, as informed by behavioral economics (Thaler, 2000), is not a realistic assumption. Instead, regulation can foster aspects that contribute to trust. Hill and O'Hara (2006), for instance, argue that a (trusting) party has a maximum level of uncertainty and vulnerability that they are willing to accept. If that threshold is crossed, the "leap of faith" (see Section 3) cannot occur. However, if the law can reduce uncertainty and vulnerability to a level that is acceptable to the trusting party, a trust relationship can still occur, thereby impacting the trust relationship overall (Hill & O'Hara, 2006). We agree with Hill and O'Hara (2006) that regulation overall can impact the uncertainty of an interaction. However, it is unlikely to completely reduce it to a level that would fully eliminate the need for "taking a leap of faith" (Möllering, 2006). Multiple factors play into this, such as the way a regulation needs to be enforced, which not only depends on the knowledge of an individual on how to enforce it but also requires effort in terms of time and money (e.g., when a case needs to be argued in front of a court). In addition, law is rarely perfectly enforced or enforceable. These practical considerations must be accounted for when deliberating about the interplay of trust and regulation. They suggest considering how regulation can help reduce uncertainty to a level where interactions can emerge. Furthermore, yet another argument supporting the near impossibility to fully eliminate the need for taking a leap of faith comes from Shapiro (1987), who argues that the structures to enforce the law can be a source of trust but also, paradoxically, provide the means for its abuse.[5] At the core, the main reason for this opportunity for abuse is that these guardians' behavior and own judgment cannot be dictated, and alternatives such as procedural norms defining the limit of legitimate power and how to constrain authority only provide inadequate counterweight, not the least because it creates a spiral of who guards the guardians (Shapiro, 1987).

### 2.2. Law and trust: Optimization

Building on an assumed co-existence of law and trust, the normative question of the extent to which law *should* impact relationships arises, thus addressing the point of optimization. It seems only reasonable to minimize distrust or promote "taking a leap of faith" when it is warranted to do so (see also Starke & Ienca, 2022). The role of law boils thus down to a definition and understanding of what *optimizing trust relationships* means. We see that the question of the role of law in trust requires a nuanced analysis and depends strongly on context and the underlying motivation of regulators. For instance, if the motivation is to protect uneducated consumers from being sold "bogus" AI services, then such services can be banned altogether, which also sows distrust in providers of such services. In fact, regulation has different tools to work toward more competent automation overall, such as constraining or mandating certain features of automation (e.g., security). However, the focus on constraints or obligations would be too narrow, as regulation can help produce relevant information necessary to level the playing field (Gasser, 2016). As Hill and O'Hara (2006), in a key article on this very topic, explain (p. 1758): "Law also can help produce trust-relevant information. Auditing and monitoring by government and private entities can help to produce trust-relevant information for others to rely on. But promoting trust may sometimes require limiting access to information. For instance, the privacy of parties transacting online is protected with monitoring and enforcement of company privacy policies and through the imposition of rules about the use of information." Other authors, such as Hult (2018) have also acknowledged the complex role of information requirements, as too much information can lead to overload and not necessarily result in more trust in the information provider.

The question of what optimizing trust relationships means in different contexts has also been addressed from a more macro-level approach, describing different stances within the interplay of law and trust. Hall (2002) distinguishes between predicated (where trust is assumed), supportive (where trust is seen as something positive), and skeptical (where distrust is seen as something positive) stances. For instance, in liability law, it is assumed that trust existed but that due to a malpractice trust must be restored (predicated stance). In other situations, trust is seen as something to proactively support (supportive stance), for example, supporting fair procedures, or to actively undermine (skeptical stance), as trust might sometimes lead to long-term negative consequences (e.g., environmental regulation). This skeptical stance links to an aspect that has often been overlooked: Law can help parties signal the limits of trust, or, in other words, how much trust a trustor should put in the trustee. Hill and O'Hara (2006) mention the example of security alarm contracts, where parties can craft their contracts in a manner to indicate what the limits of the security systems are. Here the contractual provisions are drafted in a way to clearly decrease trust in the performance abilities of the security company, so that homeowners do not overtrust the abilities of the security company.

These features of regulatory tools lead to the assumption that regulation can promote trust with artificial agents (Hult, 2018). Hill and O'Hara (2006) even argue that regulation should seek to optimize trust, that is, reduce situations of undertrust as well as overtrust through regulation in order to alleviate socially and individually suboptimal scenarios. Such a view, though, is quite instrumental, in the sense of understanding law as an instrument to turn on the right knobs to achieve trust—something at best questionable on whether it can be achieved.

### 2.3. Law and AI: External factors to establish and support trustworthiness

The literature review illustrates that regulation can reduce the uncertainty of interactions by setting external factors to push the trusted agent to act according to the best interest of the trusting party (e.g., by enacting fiduciary duties or by punishing actions that harm a trusting party). These findings can be applied to the context of AI. Primarily, and as discussed within the policymaking discourse (e.g., HLEG, 2019), the goal of policymakers is to create a framework of incentives that promote the development of trustworthy artificial agents, meaning agents who are (extrinsically via regulation) motivated to uphold the trust placed in them (and act in the interest of the trusting agent) (see Hardin, 2002). Regulation thus impacts the interaction between the human and trusted artificial agent by reducing the uncertainty of the interaction to an acceptable degree. Especially within the context of human-machine interactions, optimizing the trust relationship is central, as both distrusting AI fully, and thereby underexploiting its potential, as well as overtrusting AI, leads to societally unwelcomed situations. Thus, regulation can be seen as the mechanism to "establish and support trustworthiness," which in turn is "the best device for creating trust" (Hardin, 2002, p. 30). Yet, regulation that establishes the parameters of how an artificial agent must interact with humans and redress mechanisms in case harm occurs never leads to full control over the interaction and thus a certain uncertainty and vulnerability will remain in almost all interactions (see Section 2.1). This is also why the interplay of regulation and trust needs to turn to the question of optimization. To understand the ways regulation can optimize the trust relationships among humans and machines, there is a need to understand which factors impact the trust relationship in the first place and then to understand how regulation can influence those factors (Section 3.2).

## 3. The complexity of trust in automation

### 3.1. Conditions of trust

Trust involves a relationship between at least two agents with respect to a specific action or inaction (so-called three-part relationship: Hardin, 2002). As the term "agent" indicates, a certain level of freedom of action or inaction is necessary for trust to evolve (Keymolen, 2016). The action or inaction can also be understood more broadly as a domain of activities that the first agent trusts the second agent to take care of, which can be linked to the social role a particular agent has in a specific situation (Meyerson et al., 1996). Tied to these roles are institutional arrangements that provide formal structures and enable trusting unknown others (Bodó, 2021; Zucker, 1986). Institutionalized trust, or system trust, refers to this broader trust relationship with abstract systems and has been a topic of research especially in political science (Fukuyama, 1996; Sztompka, 1999). Like trust between humans, trusting systems or institutions also requires a learning process (Keymolen, 2016).

Moreover, the rise of increased connectivity has led to research on mediated and distributed trust (Bodó, 2021; Botsman, 2017; Keymolen, 2016) as well as analyzing human-machine trust relationships (Hoff & Bashir, 2015; Lee & See, 2004; Lewis et al., 2018; Yang & Wibowo, 2022). Different authors have aimed at disentangling various forms of trust on the spectrum of more relational (deep, thick, community-based, identification-based) trust relationships on the one end, and more rational (calculative, shallow, thin, economic) trust relationships on the other. Relational trust relationships are based on norms of communal exchange, a positive emotional connection between the parties (i.e., intrinsic interest and concern for the wellbeing of the other party/parties), and a networked and broad range of interactions among the parties (Hardin, 2002; Nooteboom & Six, 2003; Sheppard & Sherman, 1998). Here, trust develops gradually and grows over time with repeated, mutual interactions (Hardin, 2002; Möllering, 2006; Nooteboom & Six, 2003). On the other hand, rational trust relationships focus on knowledge (Rousseau et al., 1998).

For a trust relationship to emerge, there must be *uncertainty* that surrounds a specific action or inaction and *vulnerability* of the trusting party. Both are reasons why a certain "leap of faith" (Möllering, 2006, p. 105, 110) is necessary as this enables the uncertain position of the trusting party to be overcome and leads the trusting party to accept the resulting exposure and potential harm (vulnerability) (Deutsch, 1958). However, uncertainty should be understood as a spectrum too, and for a trust relationship to occur we need a *lack of qualified predictability*. Thus, uncertainty must cross a certain threshold that becomes (subjectively) qualified for the trusting party and leads to vulnerability. *Vulnerability* indicates that for the trusting party something meaningful must be at stake (Mayer et al., 1995; Rousseau et al., 1998) and that the trusting party is accepting to be in this state of vulnerability (Keymolen, 2016). In a trust relationship, this vulnerability is accepted (Rousseau et al., 1998), meaning that the trusting party does not only show a "willingness to be vulnerable" but also acts upon it (Mayer et al., 1995, p. 724). Mayer et al. (1995) distinguish here between trust (the mental acceptance or willingness to assume a risk) and "risk-taking in a relationship" (actually assuming the risk of vulnerability) (p. 715), which they see as the outcome of trust as a cognitive process. Lastly, trust leads to a confident expectation (or faith) put in the trusted party. Many authors have described this condition of securely and assertively expecting the wanted outcome of an interaction to occur, with Botsman (2017) defining trust accordingly as "a confident relationship with the unknown" (p. 20).

### 3.2. Factors impacting the trust relationship

#### *3.2.1. Propositions to analyze the interplay of trust and regulation*

The following framing and discussion of the propositions results from an iterative process conducted within a multidisciplinary team with expertise in law, political science, computer science, and sociology. We analyzed and discussed a wide range of literature, worked with examples, and continuously integrated different disciplinary viewpoints. Although the article is conceptual, we were loosely guided by grounded theory (Strauss & Corbin, 1994), especially as our approach is inductive and resulted in the description of hypotheses (i.e., propositions) which can be tested in future research. Anchored in this, we first leveraged the literature review on trust and the three-part relationship (Section 3.1) to derive the factors that influence trust into three areas: The ones that belong to the sphere of the trusting party (the human), the ones that belong to the sphere of the trusted party (e.g., the artificial agent), and the ones that belong to the sphere of the trusted act. We then added onto this construct all factors pertaining to the environment in which the three-part relationship took place (see Fig. 1). We analyzed the literature on trust and automation and extracted the relevant factors that were discussed. In the following, we refer to these factors as *propositions*: Despite existing research, there remains a need for empirical testing of the influence of regulation on trust through each of these factors. We back the propositions and their classification with literature from the fields of human-automation trust, and, in line with a contemporary tendency toward the alignment of the discussions on the human-automation trust relationship, with frameworks of human-human trust features (Lewandowsky et al., 2000; Madhavan & Wiegmann, 2007; see here also studies discussing the measurement of trust in automation such as Jian et al., 2000; Madsen & Gregor, 2000; Yagoda & Gillan, 2012). The analogy between human-automation and human-human trust models is not surprising, especially when considering research that shows the intricate and socially constructed relationships humans create with technology (Epley et al., 2007; Guzman, 2018; Nass et al., 1994, 1996, 1999).

*3.2.1.1. Proposition 1: The legal and institutional frameworks impact the trust relationship.* Legal and institutional frameworks are only one part of what constitutes the law. Even with statutes and institutions in place, there needs to be policy decisions on how to enforce the law, decisions which range to resources put in place to detect non-compliance, to which incentives and disincentive to have to nudge or deter, to how to raise awareness of these policy choices (as deterrence acutely rests on perception) (Gibbs, 1985). With this distinction in mind, intuitively, regulation immediately impacts the structure in which human-automation relationships emerge. The legal and institutional frameworks in which a relationship emerges provide "structural assurances and situational

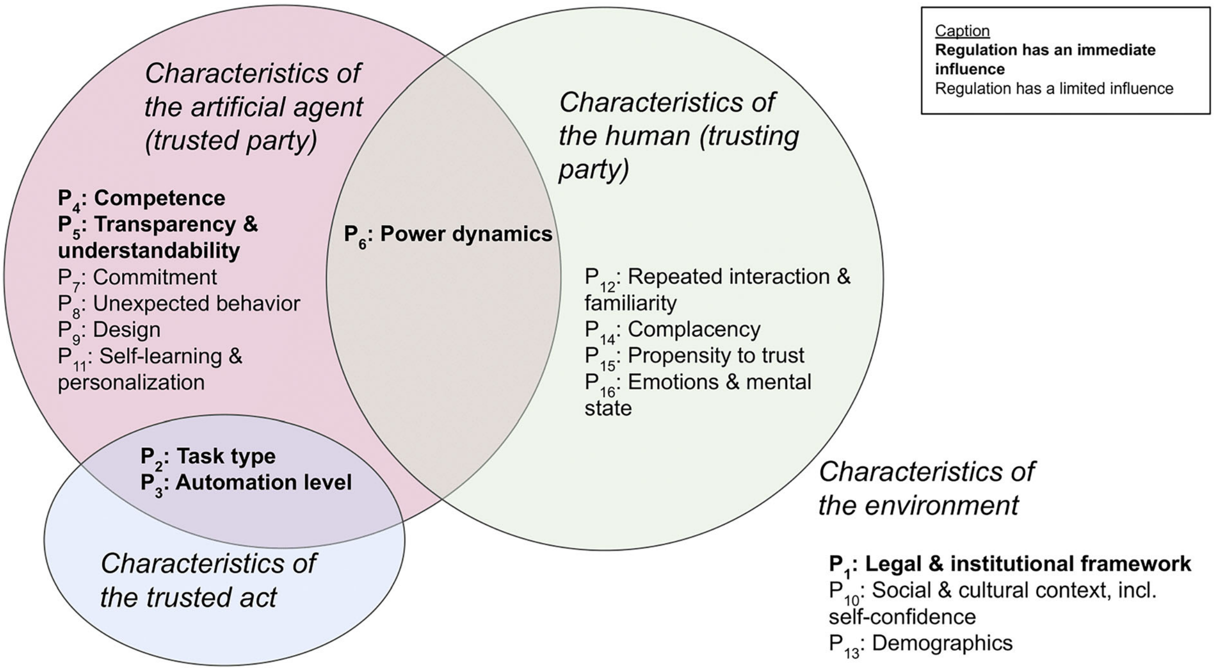

Caption
Regulation has an immediate influence
Regulation has a limited influence
Characteristics of the artificial agent (trusted party)
$P_4$: Competence
$P_5$: Transparency & understandability
$P_7$: Commitment
$P_8$: Unexpected behavior
$P_9$: Design
$P_{11}$: Self-learning & personalization
$P_6$: Power dynamics
Characteristics of the human (trusting party)
$P_{12}$: Repeated interaction & familiarity
$P_{14}$: Complacency
$P_{15}$: Propensity to trust
$P_{16}$: Emotions & mental state
$P_2$: Task type
$P_3$: Automation level
Characteristics of the trusted act
Characteristics of the environment
$P_1$: Legal & institutional framework
$P_{10}$: Social & cultural context, incl. self-confidence
$P_{13}$: Demographics

**Figure 1** Propositions (P) and their influence on the human-automation trust relationship.

normality, produce common knowledge and shared expectations, and provide various forms of safeguards and guarantees against uncertainties." (Bodó, 2021, p. 2670). A key aspect is the (perceived) stability and predictability that legal and institutional environments provide when enforcement mechanisms that uphold agreed-upon norms are in place, leading to uniform practices (DiMaggio & Powell, 1983). Such mechanisms lead to more accountable, predictable behavior, encouraging "a more open, trustful attitude because it provides the truster with a kind of insurance against possible losses, a backup option against potential breaches of trust." (Sztompka, 1999, p. 88). It thus leads to what Hill and O'Hara (2006) call "trust-that-trust," where the legal environment and social norms incentivize reducing the uncertainty of an interaction. With respect to regulation of data processing and new technologies, research has shown that AI regulation and governance leads to more data sharing online (Chatterjee & Sreenivasulu, 2023). As the relation of a trustor to institutions and/or legal framework can be more abstract than, for instance, a specific task performed by an artificial agent, specific other factors might also play a role in this fundamentally subjective judgment: personal proclivity to trust, political inclination to authority (see also Davidovitz & Cohen, 2022), emotional stability.

*3.2.1.2. Proposition 2: The task type impacts the trust relationship.* Lee (2018) illustrates how tasks requiring different skills—from more human, and emotional, skills to mechanical ones—impact the perception of a counterparty interacting with either a human or artificial decision-maker in terms of perceived fairness, trust, and emotional response. Participants trusted both human and artificial agents equally with respect to tasks involving purely mechanical skills. However, human agents were perceived as fairer decision-makers for tasks involving predominantly human skills, even when the decisions taken by both types of agents were the same. Glikson and Woolley (2020), in their in-depth overview of factors affecting trust in AI, also discuss the role of task characteristics: They note that trust in automated systems is linked to whether the allocated task is perceived to lie within the system's actual abilities. Similarly, while not explicitly focused on trust, Hidalgo et al. (2021) show across many scenario-based experiments how evaluating the ethics of intelligent machines versus humans varies substantially based on the task characteristics and task context. For functional and mechanical tasks, machines are often preferred over (more trusted than) humans, while for social and emotional tasks the opposite is true, and individuals tend to prefer (and trust) human interaction partners (Glikson & Woolley, 2020; Im et al., 2023).

*3.2.1.3. Proposition 3: The level of automation, and the extent to which a person can adapt the level of automation impacts the trust relationship.* The level of automation, typically understood as the level of control and possibility to interfere in an automated system, impacts the human-automation-trust relationship (Schaefer et al., 2016). Typically, a positive perception of not the absolute level of automation but of its *adequacy* in a given context will lead to greater trust (Schaefer et al., 2016). In their review article on trust in AI, Glikson and Woolley (2020) identify "machine intelligence," describing system's levels of automation, agency, and autonomy as an important factor in fostering trust. This is connected to user expectations and perceptions of adequacy. For example, in the context of embodied AI, "to trust and accept a robot's actions, the task allocated to the robot should be well matched to its actual abilities" (p. 635). In a different context, a TSA officer might be trustful of an x-ray machine automatically detecting the presence of weapons when the luggage is empty but less so when it is full of other possibly ambiguous objects (Merritt et al., 2013). Moreover, the ability of an individual to adapt the level of automation (notably to exert control) increases trust in the overall system (Sanders et al., 2014).

*3.2.1.4. Proposition 4: The (perceived) competence of an artificial agent impacts the trust relationship.* In human-automation trust literature, central systemic properties are performance and reliability (Hancock et al., 2011; Madhavan & Wiegmann, 2007; Muir & Moray, 1996), which includes aspects such as operational safety and data security (Hengstler et al., 2016). Highly reliable systems are ones that competently perform and thereby engender trust with respect to a given task (Lewis et al., 2018; Schaefer et al., 2016). The relationship between reliability, trust, and reliance was especially analyzed in the context of autonomous vehicles (see e.g., Bliss & Acton, 2003; Yamada & Kuchar, 2006). However, it is difficult to extrapolate findings in one study on perceived performance and reliability of technology to other contexts, due to "the instability of the technology and the fragmentation of contexts and experiences" (Bodó, 2021, p. 2681).

The literature on (perceived) competence has also focused on erroneous behaviors of automation and its impact on the overall trust relationship. Much focus has been put on trust repair (when errors occur). As expected, studies show that overall trust declines when errors occur (Lee & Moray, 1992; Lee & See, 2004; Muir & Moray, 1996). However, many factors influence said trust declines, such as personal properties (e.g., self-confidence) and relational properties (e.g., performance reliability) (Lee & Moray, 1992). Moreover, the easier the task that the automation system fails at, the higher the decline in trust (Madhavan & Wiegmann, 2007), which could be tied to findings that indicate that humans accept fewer errors from automated systems (and holding them to a higher accuracy standard than humans) (Madhavan & Wiegmann, 2007; Wiegmann et al., 2001) as well as findings that indicate that we not only pay more attention to errors of automated systems but remember their errors longer (Dzindolet et al., 2002). The literature, therefore, discusses the processes of adjusting trust over time, not only to past experiences (e.g., breaches) but changes of circumstances (so-called trust calibration), as well as trust repair, that is, ways to reinstate trust in cases of breaches, for example by determining the cause for the breach and setting in place formal structures to prevent future negative events (e.g., legal mechanisms such as contractual norms or social mechanisms such as out casting) (see e.g., Gillespie & Dietz, 2009).

*3.2.1.5. Proposition 5: Transparency and understandability impact the trust relationship.* There is evidence that transparency enhances trust in automated systems (Glass et al., 2008) and overall confidence in a system (e.g., in recommender systems: Pu & Chen, 2007). However, these results might not be reproducible across different contexts (and especially not when deployed in the real world; see also Felzmann et al., 2019, 2020 for the relation between transparency of AI systems and trust). In addition, the better and the more accurate the understanding of a human operator is vis-à-vis the functioning of an automated aid, the greater is the trust placed in the automated aid for a given task that the human operator deems within the range of actions of the human aid (Dzindolet et al., 2003; Wang et al., 2016). Understandability is not only important for placing trust, but also for better understanding why errors occur (Dzindolet et al., 2003), potentially leading to a faster re-calibration of trust instead of creating overall distrust in the system. With respect to machine learning-driven systems, research has shown that explanation-driven interactive systems lead to greater user satisfaction (Guo et al., 2022).

Related to the topic of transparency and understandability is the opaqueness of the intentionality of artificial agents. Unlike in most human-to-human relationships, technology itself does not "feel" the consequences of a breach of trust. Therefore, the human-automation relationship has been characterized as lacking intentionality (Lee & See, 2004). The lack of intentionality is then also tied to larger discussions on how to assign responsibility and blame in instances where trust was breached (Cheshire, 2011) with humans often being quick to blame technology for errors, yet more reluctant to attribute positive outcomes to the artificial agents (Madhavan & Wiegmann, 2007).

*3.2.1.6. Proposition 6: Power dynamics impact the trust relationship.* Power dynamics, including information and power asymmetries among a human agent and an artificial agent and the entity that manages the artificial agent impact the trust relationship. Several scholars have argued how the power of companies, and the way that they frame debates in their own favor, is inherently asymmetrical as customers struggle to question companies' actions and motivations (Van Dijck et al., 2018). This, in turn, impacts the trust of users into artificial agents manufactured by these companies (Nowotny, 2021). More specifically on artificial agents, apart from the task per se being assigned to the agent, the power-trust dynamic has remained understudied in empirical terms.

*3.2.1.7. Proposition 7: The commitment of an artificial agent impacts the trust relationship.* Commitment includes the notion of benevolence, fiduciary responsibility, the duty of care, and integrity. Benevolence describes the commitment with respect to care and concern for the trusting party's interests and well-being (Bodó, 2021; Mayer et al., 1995; Searle et al., 2011). It hence leads to the incorporation of the trusting party's interests by the trusted party. Integrity refers to the adherence to moral, ethical, and legal principles which build the basis for established codes of behavior within a given community (Bodó, 2021; Mayer et al., 1995). Within the AI literature, much focus has rested on the *alignment problem* (Christian, 2020; Gabriel, 2020), that is, on studying whether an artificial agent's goals or values align with those of a human. There are multiple ways of designing how artificial agents align with human goals and values: from clear instructions, to capturing revealed preferences, to augmenting those revealed preferences with ones that a human would have under rational circumstances, or even aligning those with overarching moral beliefs of a community (Gabriel, 2020). However, often, human goals and values are fluid, context-dependent, and change over time. To remedy this situation, AI researchers have proposed that the uncertainty about goals and values could be integrated in the formalization of the objective function (Russell, 2020). This results in an artificial agent that will ask human operators for additional specification of the objective function at run time to be able to determine its value. An artificial agent that acts upon unknown preferences in this way could act more benevolent toward the human as it can take shifting human goals and values into account by involving the human directly.

*3.2.1.8. Proposition 8: Unexpected behavior impacts the trust relationship.* An artificial agent might produce an unexpected (yet correct) answer (Lyons et al., 2023). The lack of predictability as well as the often emotive aspect of trust is why we have different reactions when an artificial agent behaves unexpectedly, yet without it being erroneous behavior (see Proposition 4 on competence for this matter). Nonetheless, even (correct) unexpected behavior tends to be mitigated with trust-repair-strategies discussed within the literature on underperforming agents (Kox et al., 2021), such as providing explanations for why a specific unexpected behavior occurred (Lyons et al., 2023). In that sense, the unexpected behavior of artificial agents is perceived similarly to unreliable behavior.

*3.2.1.9. Proposition 9: Design impacts the trust relationship.* Automation can take different shapes and forms, which impacts the human-automation trust relationship (Glikson & Woolley, 2020); these forms can be from purely virtual to embodied systems. In particular, physical presence (Bainbridge et al., 2011), but also visual representations of artificial agents in purely virtual settings, tend to increase the trust placed in the artificial agent (Glikson & Woolley, 2020). Aesthetical appearance likewise leads to more likable artificial agents (Schaefer et al., 2016), typically also by creating a more anthropomorphized experience (Schaefer et al., 2016; Waytz et al., 2014), while being aware of uncanny valley effects (Bartneck et al., 2009). Importantly though, the physical appearance and form should be aligned with the functionalities the artificial agent has to perform and thus meet the expectations individuals have with respect to the capabilities of the agent (Schaefer et al., 2016).

*3.2.1.10. Proposition 10: The social and cultural context impacts the trust relationship.* The context in which automation is deployed includes social and cultural aspects (e.g., the definition of social roles and expectations attached to them, cultural acceptability of certain behavior) as well as how society generally perceives said technology. For instance, with respect to AI, there is much hype and fear (Cave & Dihal, 2019), and marketing campaigns that sell AI as "magic" do not help individuals to have an accurate mental depiction of the technologies' capabilities and limitations (Knowles & Richards, 2021, with reference to Elish & Boyd, 2018). These narratives create expectations, and frustration when the expectations are not met (Glass et al., 2008). Thus, communication about a new technology has an important impact on the social reactions it triggers, and thereby on the diffusion rate of new technologies (Hengstler et al., 2016). However, there is just so much communication can achieve, as cultural factors and differences also impact overall trust in automation, as much as all the other factors listed in this paper interacting as part of a complex mesh. Measuring such cultural differences is a challenging task, with various authors having proposed and validated different approaches (e.g., Chien, 2016).

*3.2.1.11. Proposition 11: Self-learning abilities, including the ability to customize and personalize behaviors, of artificial agents impact the trust relationship.* An important aspect that has not yet been analyzed in depth is what happens to the human-automation trust relationship when interacting with machines that learn over time and adapt their behavior to the party they are interacting with (personalized interactions). In fact, the learning capabilities of newer artificial agents might not only challenge the trust establishment overall: Learning capabilities might impact the tolerance of users concerning unexpected or even erroneous behavior (similarly individuals might sometimes "forgive" Alexa for a mistake as the system is still learning), yet this proposition remains to be tested. What has been shown in the literature is the existence of a so-called personalization paradox, where greater personalization increases the relevance of a service, and yet also increases an individual's sense of vulnerability and thus leads to a decrease in adoption (Aguirre et al., 2015).

*3.2.1.12. Proposition 12: Repeated interactions and familiarity impact the trust relationship.* Rempel et al. (1985) suggest that "as relationships progress there is an inevitable shift in focus away from assessments involving specific behaviors, to an evaluation of the qualities and characteristics attributed to the partner" (p. 96). While in a human-to-human interaction, such an argument is intuitively relatable, various studies have similarly shown the impact of familiarity through repeated interactions on trust (Glikson & Woolley, 2020). However, findings as to how exactly this impact could be characterized have been inconsistent. For instance, studies on robotic AI indicated an initial low level of trust that develops over time toward higher levels of trust, while interactions with virtual agents or bots showed the opposite trust trajectory, with high initial trust that decreases over time (Glikson & Woolley, 2020). Important for this proposition remains the evolution of the trust relation predicated on repeated interactions which links to the topic of familiarity (Yang & Wibowo, 2022). Familiarity and prior experiences with artificial agents can impact the trust relationship in either way: While a positive past interaction facilitates trust, a negative one typically reduces trust in the artificial agent (Yang & Wibowo, 2022).

*3.2.1.13. Proposition 13: Demographics (age, gender, socio-economic status) impact the trust relationship.* Research on age as a factor affecting trust is limited, focusing on a small number of participants and often on autonomous vehicles (Donmez et al., 2006). The studies within these fields suggest that younger and middle-aged adults are less likely to trust automation, while older adults are more likely to trust automation. However, it is difficult to generalize the results of the relevancy of the impact of individual traits (e.g., age) across domains (Schaefer et al., 2016). Specifically, in relation to AI (rather than automation more generally), the directionality is the opposite, with younger people trusting such technology more strongly (Gillath et al., 2021). This is summarized in a recent article (Yang & Wibowo, 2022) that discusses how demographic characteristics, including age, gender, and socio-economic status, affect trust in AI. For gender, and across several studies reviewed, men are more trusting toward AI than women (Yang & Wibowo, 2022). Gender overall has a stronger impact than age and socio-economic status, and gender differences can be explained in relation to socialization and privacy concerns (Shao et al., 2020).

Finally, socio-economic status in most studies is captured by the respondents' education level, which is positively related to trust in AI. An explanation for this is "that higher education provides people with more knowledge of risk management […], which may reduce their aversion to the risks of adopting new technologies" (Yang & Wibowo, 2022, p. 14).

*3.2.1.14. Proposition 14: Complacency and automation bias impact the trust relationship.* Another relevant aspect is the role of complacency (Bagheri & Jamieson, 2004; Parasuraman & Manzey, 2010). Complacency refers to human agents not paying enough attention to the actions of an artificial agent even if it would be advisable to do so (e.g., not paying attention to an autopilot, Parasuraman & Manzey, 2010). Yet, as with other aspects, complacency arises due to multiple factors, such as situations in which individuals are multitasking (Molloy & Parasuraman, 1996), or due to an initial attitude toward automation (Parasuraman & Manzey, 2010, p. 389). A person's attitude toward an automation system is further influenced by automation bias, that is, the tendency of individuals to ascribe greater importance or correctness to automated systems (Parasuraman & Manzey, 2010). A factor influencing automation bias is the "phenomenon of diffusion of responsibility" (Parasuraman & Manzey, 2010, p. 392) between human and automated systems. Thus, the task context and who is made accountable for a decision impacts the overall level of complacency.

Complacency and automation bias "share several commonalities" (Parasuraman & Manzey, 2010, p. 397). They both are a kind of automation misuse. Related to automation bias and complacency is the misuse of technology due to overtrust. Overtrust means that "an operator overestimates the actual reliability of an aid" (Madhavan & Wiegmann, 2007, p. 281). Overtrust, however, has a more active meaning (especially than complacency) where an individual (intentionally) takes more risks because it trusts that the system will reduce the risk of something going wrong (Itoh, 2012). Overtrust has been studied in the context of robot aids (Gaudiello et al., 2016; Robinette et al., 2016; Salem et al., 2015), exoskeletons (Borenstein et al., 2018), self-driving cars (with a driving robot) (Kundinger et al., 2019), and a research agenda to tackle the problem of overtrust has been proposed (Aroyo et al., 2021). While overtrust can occur because of a miscalculation of how well the automated system works (e.g., reduces the risks of an accident), it can also result from a misunderstanding of what the automated system can in fact do (Borenstein et al., 2018; Itoh, 2012).

*3.2.1.15. Proposition 15: The propensity to trust impacts the trust relationship.* Trust propensity is a dispositional willingness to trust another agent (Colquitt et al., 2007; Gill et al., 2005; Mayer et al., 1995). For instance, individuals who are insecure about their own ability to perform a task tend to place greater trust in automated aids

(Hoff & Bashir, 2015), while greater self-confidence tends to decrease trust placed in external (human or automated) aids (Lee & Moray, 1994; Madhavan & Wiegmann, 2007). The propensity to trust is also shaped by past experiences (Searle et al., 2011). Yet, one's propensity to trust might more strongly influence trusting beliefs and intentions and ultimately enabling "risk-taking behavior" (as framed by Mayer et al., 1995) than past experiences (Colquitt et al., 2007; Govier, 1994).

*3.2.1.16. Proposition 16: Emotions and mental states impact the trust relationship.* Emotions affect the relationship with automation/technology and reliance on automation. Overall, happiness has been shown to increase trust in automation, especially during the first interaction with it, as well as the liking (defined as the extent to which an individual feels positively toward a technology) of the technology one interacts with (Merritt, 2011; Merritt et al., 2013). Liking or positive attitudes are shaped by different emotive factors, such as the level of comfort which is tied to how familiar and close an individual is with an artificial agent (Schaefer, 2013). Likewise, mental states (e.g., fatigue, stress) impact the relationship and interaction with others as well as automation (Schaefer et al., 2016). With respect to automation, low attentional control has been shown to lead to higher reliance on automated systems (Schaefer et al., 2016).

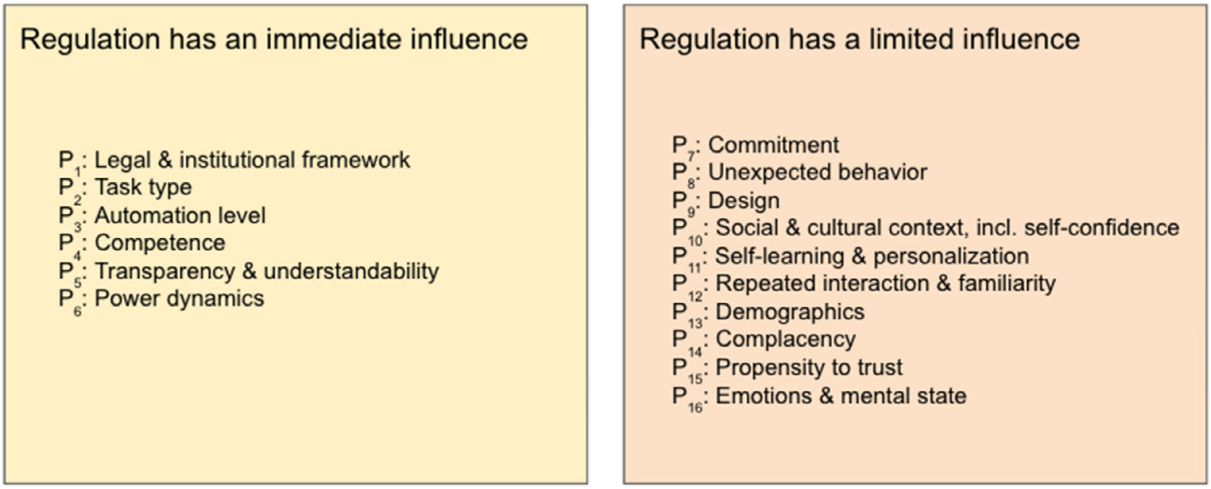
### *3.2.2. Future empirical research*

Figure 1 provides an overview of the propositions that are explained in more detail above. These propositions are inter-related and can be attributed to different spheres, such as ones related to the characteristics of the artificial agent (trusted party) (e.g., competence, transparency, commitment, unexpected behavior, design), characteristics of the human (trusting party) (e.g., repeated interactions and familiarity, complacency, propensity to trust, emotions and mental states), characteristics at the intersection of the trusted party and trusting party (e.g., power dynamics), characteristics of the trusted act (e.g. type of task, automation level), and characteristics of the environment (e.g., legal and institutional framework, social and cultural context, demographics). The categories are based on the literature review in Section 3.1, and as shown, the interactions between them are highly complex and bi-directional.

While the propositions are backed by interdisciplinary literature, they remain hypotheses that require empirical testing to determine their validity. Each set of propositions might require a different methodology, ranging from more qualitative to quantitative measures. For instance, to understand the impact of the legal and institutional framework (Proposition 1) requires a comparative approach between different jurisdictions, while other propositions such as the impact of the task (Proposition 2) or level of automation (Proposition 3) could be tested in experimental studies by keeping all other factors constant and only varying variables such as the task type or level of automation, respectively. Qualitative research, for example through interviews and case studies, would allow the exploration of the meanings people attach to the different factors affecting human-automation trust, offering contextual nuance. For example, when it comes to the transparency/understandability-trust link (Proposition 5), the think-aloud technique (Nielsen et al., 2002) could elicit rich descriptions of how people perceive a specific AI-related technology, such as a chatbot, in terms of its transparency, and how such transparency perceptions may foster or reduce trust. For testing some propositions, behavioral data, for example from digital platforms, will yield fruitful results. Proposition 8 about unexpected behavior and Proposition 9 about design, for instance, can be tested with natural or field experiments, answering questions such as the following: Does unexpected behavior of an AI service (e.g., when changing its business model abruptly) diminish user trust? How do design changes, for example considerably altering the weights of a social media algorithm, affect trust levels?

### *3.2.3. Impact of regulation on the propositions*

The wealth of propositions impacting trust relationships leads us to consider to what extent regulation could genuinely affect all or some of them, and to which degree. We posit that we can distinguish between propositions where we see an *immediate influence* on elements that impact the human-automation trust relationships and ones where that influence is more *limited* (see Fig. 2 and the following discussion of the classification). As mentioned at the beginning under "Proposition to analyze the interplay of trust and regulation" within this conceptual article we arrived at our observations by leveraging real-world examples and discussing them within a multidisciplinary team. In addition, we ground our observations within the literature as well the analysis of specific provisions within laws that govern automation and challenges that arose from increased data processing.



Regulation has an immediate influence

$P_1$: Legal & institutional framework
$P_2$: Task type
$P_3$: Automation level
$P_4$: Competence
$P_5$: Transparency & understandability
$P_6$: Power dynamics

Regulation has a limited influence

$P_7$: Commitment
$P_8$: Unexpected behavior
$P_9$: Design
$P_{10}$: Social & cultural context, incl. self-confidence
$P_{11}$: Self-learning & personalization
$P_{12}$: Repeated interaction & familiarity
$P_{13}$: Demographics
$P_{14}$: Complacency
$P_{15}$: Propensity to trust
$P_{16}$: Emotions & mental state

**Figure 2** Propositions (P) and the impact on regulation on each of them.

However, we note that there is a lack of literature on how regulation in particular impacts the human-automation trust relationships. Considering this knowledge gap, more empirical studies in this field are needed. Our article provides a foundation for future empirical research analyzing the influence of regulation on different factors that impact trust relationships. Our approach extends the knowhow of the field of trust in automation to the regulatory field by providing a framework to discuss certain regulatory mechanisms. While Section 3 provides the theoretical background, we apply our analysis within Section 4 to the AI Act of the EU.

*3.2.3.1. Immediate influence of regulation on propositions P1–P6.* The legal and institutional framework (Proposition 1), type of task (Proposition 2), level of automation (Proposition 3), perceived competence of the artificial agent (Proposition 4), transparency and understandability (Proposition 5), and power dynamics (Proposition 6) impact the trust relationship between a human and an artificial agent and are aspects that regulation can shape in a direct manner. In democracies, the legal framework can be amended to address novel issues (e.g., proposed AI Act, see Section 4) and the institutional setting changed according to the needs of society. Regulation thus lays down the way in which interactions occur and determines the underlying *characteristics of the environment* in which trust relationships emerge.

With respect to the *characteristics of the trusted act* itself, regulation can directly mandate which acts are allowed and if so to what extent and in what form. For instance, the use of real-time biometric identification in public spaces for law enforcement purposes can be generally banned and only allowed under specific circumstances (Art. 5 (1) (d) AI Act) or the application of automated decision-making in the law enforcement context only allowed if certain appropriate safeguards are in place (Art. 11 Law Enforcement Directive). Thereby the types of tasks that can be left to automation are directly shaped by regulation (Proposition 2). Likewise, and linked to the types of tasks, the level of automation can be regulated (Proposition 3). For instance, the General Data Protection Regulation provides for a right (or prohibition, depending on the reading, see Bayamlıoğlu, 2022; Tamò-Larrieux, 2021) to obtain further information on automated decision-making systems that have a significant impact on individuals. Such a norm impacts the level of automation that companies employ.

With respect to the *characteristics of the artificial agent* (trusted party), regulation has an immediate influence on the competence of a system as it can oblige developers of AI systems to disclose performance-related information about the system, thus affecting perceived competence and trust (Proposition 4). Regulation can also prohibit undue advertising and marketing that inflates the competence of a system, and it can impose standards and safeguards. Especially when a system is considered high risk, more scrutiny is put on the overall reliability and competence of such a system (as the AI Act exemplifies, see Section 4). Moreover, and also exemplified in the AI Act, regulation can immediately influence transparency through information disclosure requirements (Proposition 5). We have seen this also in other regulations, such as the Digital Services Act, that contains for instance transparency obligations for online platforms with respect to the measures they implement to moderate online content

(Art. 24 Digital Services Act). Different forms of information requirements depending on the target audience could be required by law, thereby also influencing the understandability of how an artificial agent operates.

At the *interaction-point* of the trusting and trusted party, we posit that power relations become relevant: This is particularly important, when, as apparent in the technology sector, the technology providing party has much more understanding and knowhow over its product and services compared to the trusting individual. However, power dynamics can also be shaped by regulation (Proposition 6) as regulation can level the playing field (e.g., with respect to information provision), or restrict powerful market players in certain aspects. Even if regulatory attempts can lag behind and be exploited by powerful companies to their advantage (Cohen, 2019), we have seen in recent years attempts by the EU to target powerful market players (e.g., with the Digital Market Act and the Digital Services Act).

*3.2.3.2. Limited influence of regulation on propositions P7–P16.* Regulation may set standards (via governance or thresholds) for benevolence (Proposition 7), unexpected behaviors (Proposition 8), or design (Proposition 9). For example, regulation could mandate anyone releasing AI tools to seek approval via an institution in a similar fashion that, within the pharmaceutical industry, a new drug product would have to go through a review process with a regulator before going to the market. For benevolence, this would mean imposing fiduciary duties and duty of loyalty (Balkin, 2020; Richards & Hartzog, 2021); for unexpected behavior showing that the frequency of such unexpected behavior is within a certain acceptable range; and for design, putting limits, for instance on how similar to humans they could look like. Yet, there are important limitations on all three with respect to how this could impact trust: enforcing fiduciary duties can be difficult to implement (and at a high cost, e.g., via audit trails); completely ruling out—technically or legally—unexpected behaviors is not doable and even ensuring that they remain within a certain frequency is speculative due to the very nature of such behaviors being *unexpected*; and remaining within the example of a ban on human-like design (humanoid artificial agents), anthropomorphism would still take place as humans tend to establish social relationships with devices not only because of their anthropomorphic design but also due to their underlying emotional and social needs (Epley et al., 2007; Wan & Chen, 2021). Regulating on these aspects, while within the realm of legislators' approach to law (unlike for Proposition 12–Proposition 16), would hence only impact trust limitedly.

A similar argument applies for the socio-cultural context (Proposition 10). Regulation is not only shaped by the cultural context, but we do acknowledge that it also shapes the social and cultural environment (Shapiro, 2011). In this sense, regulation on the social and cultural factors that impact the trust relationship has a certain influence. For instance, trust will emerge differently in high-crime areas than in safer ones (Nix, 2017) and specific policies can have an impact on crime (Malone, 2010). Hence, regulation can change the socio-cultural context and trust. However, this influence is less immediate and more intricate than for other factors, as it would be extremely difficult to causally link changes in socio-cultural contexts to regulation regarding the high number of interwoven factors at play. Furthermore, social and cultural factors influence the characteristics of the human's (trusting party) propensity to trust, their emotions, and mental state. Yet again, causally linking the impact of regulation on socio-cultural factors which would in turn impact personal factors can appear a futile exercise as too many factors would play a role.

We argue that regulation could influence and constrain developments toward self-learning and personalization (Proposition 11) as well as toward repeated interactions (Proposition 12), but that this is an unlikely development *at the moment* as it would be both too sweeping (and hence too much limiting future development) and too specific. Current legislation within the EU seeking to limit the use of machine learning has not made this step, and to the best of our knowledge around the world, most legislations have been trying to enable the development and use of such self-learning abilities, rather than trying to put an abrupt and all-encompassing break onto them. Similarly, states could mandate—but only to some limited extent—that certain professionals are required to use automated systems on a regular basis in order to increase trust of the professionals and hence indirectly of the public toward the automated system. Repeated interactions, more generally and for lay people, could also come in the form of other incentives, with economic rewards. Again, *at the moment*, there is no signal that there are interests in states having interests in such regulation. But we recognize that this could evolve for Proposition 11 and Proposition 12: regulators could have an interest in both limiting the self-learning abilities for

**Table 1** Examples from the AI Act on how the propositions are intended to impact the trust relationship.

| Proposition and *influence of regulation* | Non-exhaustive examples from the AI Act |
|---|---|
| P2 type of task, *immediate* | The AI Act prohibits specific tasks that could be performed with AI technologies such as social scoring (Art. 5(1)(c)). While the deployment of "subliminal techniques beyond a person's consciousness" are prohibited, this prohibition does "not apply to AI systems intended to be used for approved therapeutic purposes on the basis of specific, informed consent of the individuals that are exposed to them" (Art. 5(1)(a)). This indicates that specific types of tasks can be narrowly tailored through regulation. Likewise, the AI technologies used to assess the risk of reoffending of natural persons are forbidden (Art. 5(1)(d a)). |
| P3 level of automation, *immediate* | The AI Act demands that AI systems which fall under the high-risk category are developed in a manner that still allows for human oversight (Art. 14). This means that even if such systems could perform their tasks in a fully automated fashion, the AI Act mandates a "human-machine interface tool" to effectively control the level of automation. Art. 4a (1) (a) makes this clear by demanding: "'human agency and oversight' means that AI systems shall be developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans." Art. 14 further ties the human oversight requirement to the one on AI literacy (see Art. 14(1), Proposition 5) and demands that the oversight takes into account the risks, the level of automation, and the context in which the AI is employed (Art. 14(3)). |
| P4 perceived competence, *immediate* | Overall, the AI Act mandates: "'technical robustness and safety' means that AI systems shall be developed and used in a way to minimize unintended and unexpected harm as well as being robust in case of unintended problems and being resilient against attempts to alter the use or performance of the AI system so as to allow unlawful use by malicious third parties" (Art. 4a (1) (b)). More specifically, the AI Act sets out requirements (especially for high-risk AI systems) on how to ensure the reliability of such systems. These requirements are linked, among others, to risk management systems (Art. 9), data governance rules (Art. 10, e.g., checking datasets for errors and validating them, bias detection), technical documentation requirements (Art. 11) including automatic recording of events (logs) (Art. 12), and following a by design and default notion to ensure strong accuracy, robustness, safety, and cybersecurity (Art. 15). Regarding the latter point, i.e., assurance of appropriate levels of accuracy, robustness, safety and cybersecurity, benchmarks will need to be established by the relevant authorities (e.g., AI Office, European Union Agency for Cybersecurity) (Art. 15(1 a and b)). These requirements influence operational safety and security of artificial agents and thereby their (perceived) competence. |
| P5 transparency and understandability, *immediate* | The AI Act mandates that high-risk AI systems and their operations are made transparent to end users and that these are empowered to understand the decision-making of those systems (Art. 13). To this end, Art. 4a (1) (d) defines transparency as the ability of an AI systems to allow "appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system as well as duly informing users of the capabilities and limitations of that AI system and affected persons about their rights." The AI Act further lists in Art. 13 the required information that must be provided and directly influences the transparency and understandability of an artificial agent (incl. the identity and contact details of the provider, the capabilities and limitations of the AI system, the level of accuracy, robustness, and cybersecurity (tied to Proposition 4), information about how actions by users might influence how the system performs, and necessary maintenance duties).<br>A new provision compared to the proposed AI Act is the stronger focus set on AI literacy (Art. 4b; Art. 13(3a)). The aim is to create the conditions for providers, deployers, and affected individuals to be able to understand the AI system in question. The AI literacy provision also includes measures that must be set in place by AI providers to educate their staff operating and using AI technologies. "Such literacy measures shall consist, in particular, of the teaching of basic notions and skills about AI systems and their functioning, including the different types of products and uses, their risks and benefits." (Art. 4b (3)). |

(*Continues*)

**Table 1** Continued

| Proposition and *influence of regulation* | Non-exhaustive examples from the AI Act |
|---|---|
| P6 power dynamics, *immediate* | The AI Act addresses the information and power asymmetries between users and providers of AI among others by establishing transparency requirements (Art. 13) and setting obligations for providers of high-risk AI systems such as the requirements to have a quality management system in place (Art. 17). In addition, standards will no longer solely be set by leading firms in the field, but according to Art 41, the EU agency responsible for standardization (CEN) will be a key decision-maker, hence shitting the power balance back to the state and away from companies. |
| P7 benevolence, fiduciary, duty of care, *limited* | The AI Act in part does try to influence the commitment with respect to care and concern for the trusting party's interests and well-being: For instance, under Art. 10 data governance mechanisms are set in place to ensure that possible biases within the dataset are noted, hence attempting to make the data set fairer or more representative. And yet, bias-free datasets might not exist and so, remaining skeptical of the outcome until evidence emerges is warranted (Barocas et al., 2017). |
| P8 unexpected behaviors, *limited* | Unexpected but correct behaviors will not likely be flagged within a risk assessment or when analyzing the accuracy and robustness of decision-making. In light of this, the regulatory measures that best address this proposition, at least partially, are found within the norms establishing transparency and interpretability of decisions made by artificial agents (see Proposition 5). However, the focus is rather on explaining why a behavior occurred, regardless of its "expectability." Addressing this more directly with specific provisions on flagging especially unexpected but correct behavior is yet to be established. |
| P9 design, *limited* | The AI Act does not mention the shapes or forms in which AI systems perform their tasks (e.g., purely virtual or physically embodied systems). However, regulation *could* address—but importantly, has not done so—for instance, the anthropomorphization effects created by physical social robots, especially when those take on animal or human-like features by banning human-like appearances. One can argue that the design of the AI systems, especially if further research establishes a clear link between e.g., anthropomorphization and risks for individuals, is addressed partly by requiring AI system providers to set in place appropriate data governance strategies and consider the design choices of their AI systems (Art. 10). While at this stage this article refers to design decisions with respect to the data used within a system, "design choices" could be understood more broadly to capture also decisions of the (physical) interaction design. |
| P10 social and cultural context, *limited* | The establishment of the AI Act illustrates that within the EU, policymakers early on were keen to address the risks of developments within the field of AI. Thereby, social realities and cultural perceptions played a role in the development of the overall AI Act. As aforementioned, a two-pronged impetus to initiate legislation on AI was to instill trust and to be competitive. The regulation aims hence at changing the socio-cultural context but this can only happen indirectly at best. We note that there is a mention of social benefits and activities but not in the regulation itself, only in the non-binding recitals, which gives credence to our classification of Proposition 10 as an indirect factor. |
| P11 self-learning abilities, *limited* | The adopted AI Act states in Rec. 6a: "AI systems often have machine learning capacities that allow them to adapt and perform new tasks autonomously." The classification of AI thus remains broader but is coupled to learning abilities. Thereby, learning capabilities fall within the AI Act itself, however the effect of learning behaviors when interacting with humans (e.g., in a smart home setting with social robots) is not specifically addressed within the AI Act. Yet, it can be imagined that the abilities to continue learning once put on the market *could* be restricted to specific behaviors only. |
| P12 repeated interactions and familiarity, *limited* | The AI Act does not set requirements for how often users must interact with an artificial agent, nor do repeated interactions of familiarity alter the set of rules that such agents are bound to. |
| P13 demographics, *limited* | The AI Act does neither target nor try to amend the structure of age, gender, or socio-economic status of users, nor does it set rules that differentiate between these demographic factors of a user when interacting with an artificial agent. |

(*Continues*)

**Table 1** Continued

| Proposition and *influence of regulation* | Non-exhaustive examples from the AI Act |
|---|---|
| P14 complacency and automation bias, *limited* | The AI Act does include language that makes deployers of high-risk AI systems responsible to ensure that the individuals having oversight over the AI system are made aware of automation and confirmation biases (Art. 16(1)(a b)). Yet, it remains to be seen if such educational measures will in practice reduce complacency and automation biases. |
| P15 propensity to trust, *limited* | Linked to Proposition 14, the propensity to trust is a dispositional willingness of a trusting agent to trust. The AI Act does not address this disposition as far as we can tell, and regulation overall is unlikely to be able to address in abstract terms such a highly personal characteristic. |
| P16 emotions and mental state, *limited* | The AI Act does not take into account emotions of mental states of trusting parties. In fact, regulation might at best address emotional and mental states by setting frameworks that allow for safe spaces to discuss emotional or mental states of users. |

vulgar language or abilities to recognize new faces for instance, as much as in compelling only professionals to use with care support from AI tools.

Even more limited is the impact of regulation on aspects such as demographics (Proposition 13). While in certain states such as China with its famous (but now revoked) one-child policy there has been a strict regulation to control population growth, liberal democracies have not resorted to such tools, much preferring to try to impact demographics through policies (The Economist, 2021). But it should be noted that these policies have failed to deliver the right incentive to change their birth rate, and with it their age structure (The Economist, 2021), as much as China is now struggling to reverse the effects of an aging population (Yeung, 2023). Likewise, aspects linked to personality traits, such as the complacency and automation biases (Proposition 14), the propensity to trust (Proposition 15), and emotional states (Proposition 16) are less likely to be targeted and targetable by regulation. While we can imagine certain instances where through regulation emotional states are altered (e.g., guidelines on the use of drugs for palliative care), to the best of our knowledge, these examples are cantoned to niche and limited areas.

## 4. Illustrative example: The framework applied to the AI Act

A piece of to-be-enacted regulation on AI from the EU offers the chance to illustrate the framework, including the two grades of influence (Fig. 2). In October 2017, the European Council invited the European Commission to tackle AI with a risk-based framework in order for the EU to "reaffirm its leading role in the industry" (European Council, 2017). The European Commission followed through by setting up a High-Level Expert Group in March 2018, and the group delivered its much quoted first report in April 2019 (HLEG, 2019). The report put an important emphasis on trust with 145 mentions across 41 pages. It paved the way for the regulatory proposal that the European Commission unveiled 2 years later, the proposed AI Act,[6] which underwent certain changes to accommodate generative AI following the popularity of ChatGPT (Sharma, 2023). In June 2023, the European Parliament adopted a revised version.[7] While the trilogue phase continues, already the first recitals within the European Parliament's AI Act indicate a stronger focus on "human centric and trustworthy" AI that ensures a high level of protection of EU citizens (e.g., Rec. 1, 2). Even more explicit is the new Recital 4a, stating "Given the major impact that artificial intelligence can have on society and the *need to build trust*, it is vital for artificial intelligence and its regulatory framework to be developed according to Union values enshrined in Article 2 TEU, the fundamental rights and freedoms enshrined in the Treaties, the Charter, and international human rights law" (emphasis added).

In the following Table 1, we look at the adopted AI Act (reference number: P9_TA(2023)0236) and analyze the influence of individual norms and concepts on the elaborated propositions.

A review of the current AI Act shows an accordance with our theoretical considerations developed completely independently of the Act: those propositions where regulation could have an impact on trust (Proposition 1 – Proposition 6) had clear links to specific articles; those propositions where it is more nebulous to see how regulations could impact them and in turn trust (Proposition 7 – Proposition 12)—what we termed "limited influence"—either only materialize in the Act very indirectly and when argued under a specific angle (Proposition 7, 8, 14) or not at all (Proposition 9–13, 15–16). This gives support to our review, classification, and argument of the tripartite propositions-regulation-trust relation.

## 5. Conclusion

The term "trustworthy AI" has entered the regulatory discourse (HLEG, 2019). With it, numerous discussions on how to ensure trustworthy AI and enable trusting intelligent machines have emerged. Such discussions are not surprising in light of the technological advances and shifts in perception of AI: With more and more artificial agents, such as ChatGPT and related consumer-facing AI tools, entering the public sphere, our interaction with these agents becomes more common, and concepts for describing interpersonal relationships are translated to the human-automation realm (Guzman & Lewis, 2020). Policymakers have picked up on these developments and have launched attempts to shape how artificial agents interact with people. Some of these attempts try to set the conditions for individuals to trust the technology at hand—aiming thus at holistically targeting the underlying factors that influence trust. This instrumental view of trust is dangerous, as it frames the discussion in a way to push toward trusting AI. The argument is basically: If AI is trustworthy, it would make sense to trust AI. However, these discussions omit the benefits of distrust—such as keeping away from genuinely unsafe acting (Hardré, 2016). A healthy notion of caution and skepticism against technologies that are released without proper oversight or rushed to market in emergency situations (Newlands et al., 2020) could prevent undesirable outcomes. In addition, using the word "trustworthy" in a uniform fashion neglects the fact that "trustworthy" does not have the same meaning for us all, is context-dependent as well as subject to change with new information over time. In addition, the discussion on achieving trust through regulation neglects the complexity of the topic: Not only due to the sheer number of factors discussed in Section 3 that impact the trust relationship, but also because the interplay of the factors is difficult to study. Even though many studies exist, these necessarily reduce the complexity by narrowing their scope to a few, or just a single, proposition. Möllering (2006, p. 105) likewise points out that especially empirical research on trust cannot take into account all the influences on the trust relationship "held under ceteris paribus conditions." In light of this, understanding the impact of regulation on each proposition, the interplay among each other, and how the time dimensions with the interplay in a constant evolving state, is a highly complex task that so far has not been attempted to be modeled.

Our analysis in Section 3 shows that regulation might be a better fit to impact the human-automation relationship by focusing on targeting six aspects: The characteristics of the environment, such as the overall legal and institutional framework in which an interaction is embedded (Proposition 1); the characteristics of the trusted act, such as the type of the task (Proposition 2) and the automation level (Proposition 3); the characteristics of the artificial agent (trusted party), such as the competence and reliability of the agents (Proposition 4), and the transparency and understandability of the actions (Proposition 5); and the characteristics in between the trusted party and trusting party, such as addressing the power dynamics (Proposition 6). On the other hand, we see a much more limited role of regulation in addressing further properties within the following characteristics only: the commitment of the agent (Proposition 7), the unexpected behavior or more likely the responses to it (Proposition 8), the design (Proposition 9), the socio-cultural environment (Proposition 10), the self-learning (Proposition 11), and with repeated interactions (Proposition 12), the demographics (Proposition 13), complacency (Proposition 14), a person's own propensity to trust (Proposition 15) and a person's emotions (Proposition 16).

Determining which factors are the most important presents methodological challenges (of having to repeat empirical studies many times), but, similarly and as problematic, many of the findings presented as hypotheses may find their limits when tested out repeatedly and across time (due to the changing nature of trust and of the relative importance of factors impacting it). This means that not only determining but also targeting specific factors believed to be of high importance is likely to be analytically flawed—and justifies once more why we termed these "propositions." And yet, by following up with such an approach of clearly distinguishing between factors the law can influence, regulation can become more efficiently targeted.

## Acknowledgments

Aurelia Tamò-Larrieux received funding from the International Postdoctoral Fellowship grant (Basic Research Fund, University of St. Gallen, Switzerland; project number 1031564). Clement Guitton received funding from the Hasler Foundation (project grant 21089), and Research Council of Norway (grant number 299178).

## Data availability statement

We have not used data but all our sources are cited and available on Google Scholar.

## Endnotes

[1] An idea that is not necessarily new, as for example the title of this French law indicates: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000801164/.

[2] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Civil-liability-adapting-liability-rules-to-the-digital-age-and-artificial-intelligence/public-consultation_en.

[3] We would like to point out that much of the literature has focused on automation more broadly rather than AI. Automation is broader than AI itself and can be understood as an overarching category that contains AI. At its core, AI is about using machines to perform tasks or explore domains that would otherwise require human work or assistance (including cognitive assistance). Like other forms of automation, the aim of AI in applications is to increase efficiency and consistency in handling tasks of varying difficulty. Therefore, the literature on trust and automation can serve as a baseline to describe and analyze the elements that impact the trust relationship when humans interact with artificially intelligent agents.

[4] Adopted version of the AI Act can be found here: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf.

[5] Shapiro (1987) in fact does not argue that this is the case solely for law enforcement but more widely in any context of impersonal trust with acting guardians of it defined as "a supporting social-control framework of procedural norms, organizational forms, and social-control specialists which institutionalize distrust" (p. 635).

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[7] https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

## References

Aguirre, E., Mahr, D., Grewal, D., De Ruyter, K., & Wetzels, M. (2015). Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness. *Journal of Retailing*, 91(1), 34–49.

Aroyo, A. M., De Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H., Solberg, M., & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics*, 12(1), 423–436.

Bagheri, N., & Jamieson, G. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced complacency. In D. A. Vincenzi, M. Mouloua, & P. A. Hancock (Eds.), *Human performance, situation awareness and automation: Current research and trends* (Vol. 2, pp. 54–59). Lawrence Erlbaum.

Baier, A. (1986). Trust and antitrust. *Ethics*, 96(2), 231–260.

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3(1), 41–52.

Balkin, J. M. (2020). The fiduciary model of privacy. *Harvard Law Review Forum*, 134, 11–33.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. https://fairmlbook.org/

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009). My robotic doppelgänger—A critical look at the uncanny valley. In *Proceedings of the 18th IEEE international symposium on robot and human interactive communication* (pp. 269–276). IEEE.

Bayamlıoğlu, E. (2022). The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation". *Regulation & Governance*, 16(4), 1058–1078.

Blair, M. M., & Stout, L. A. (2000). Trust, trustworthiness, and the behavioral foundations of corporate law. *University of Pennsylvania Law Review*, 149(6), 1735–1810.

Bliss, J. P., & Acton, S. A. (2003). Alarm mistrust in automobiles: How collision alarm reliability affects driving. *Applied Ergonomics*, 34(6), 499–509.

Bodó, B. (2021). Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators. *New Media & Society*, 23(9), 2668–2690.

Borenstein, J., Wagner, A. R., & Howard, A. (2018). Overtrust of pediatric health-care robots: A preliminary survey of parent perspectives. *IEEE Robotics & Automation Magazine*, 25(1), 46–54.

Botsman, R. (2017). *Who can you trust?: How technology brought us together–and why it could drive us apart*. Penguin UK.

Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1, 74–78.

Chatterjee, S., & Sreenivasulu, N. S. (2023). Impact of AI regulation and governance on online personal data sharing: From sociolegal, technology and policy perspective. *Journal of Science and Technology Policy Management*, 14(1), 157–180.

Cheshire, C. (2011). Online trust, trustworthiness, or assurance? *Daedalus*, 140(4), 49–58.

Chien, S. Y. J. (2016). The influence of cultural factors on trust in automation. Doctoral dissertation, University of Pittsburgh. https://www.proquest.com/docview/1944060499?pq-origsite=gscholar&fromopenview=true

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Cohen, J. E. (2019). *Between truth and power*. Oxford University Press.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, 92, 909–927.

Cross, F. B. (2004). Law and trust. *Georgetown Law Journal*, 93, 1457–1545.

Davidovitz, M., & Cohen, N. (2022). Alone in the campaign: Distrust in regulators and the coping of front-line workers. *Regulation & Governance*, 16(4), 1005–1021.

Deakin, S., Lane, C., & Wilkinson, F. (1994). 'Trust' or law? Towards an integrated theory of contractual relations between firms. *Journal of Law and Society*, 21(3), 329–349.

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution*, 2, 265–279.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48(2), 147–160.

Donmez, B., Boyle, L. N., Lee, J. D., & McGehee, D. V. (2006). Drivers' attitudes toward imperfect distraction mitigation strategies. *Transportation Research Part F: Traffic Psychology and Behaviour*, 9(6), 387–398.

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.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1), 79–94.

Elish, M. C., & Boyd, D. (2018). Situating methods in the magic of Big Data and AI. *Communication Monographs*, 85(1), 57–80.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.

European Council. (2017). European Council meeting (19 October 2017) – Conclusions.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6), 3333–3361.

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1), 1–14.

Fukuyama, F. (1996). *Trust: The social virtues and the creation of prosperity*. Simon and Schuster.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437.

Gasser, U. (2016). Recoding privacy law: Reflections on the future relationship among law, technology, and privacy. *Harvard Law Rev Forum*, 130, 61–70.

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

Gibbs, J. P. (1985). Deterrence theory and research. In *Nebraska symposium on motivation: The law as a behavioral instrument* (Vol. 33). University of Nebraska Press.

Gill, H., Boies, K., Finegan, J. E., & McNally, J. (2005). Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust. *Journal of Business and Psychology*, 19, 287–302.

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115, 106607.

Gillespie, N., & Dietz, G. (2009). Trust repair after an organization-level failure. *Academy of Management Review*, 34(1), 127–145.

Glass, A., McGuinness, D. L., & Wolverton, M. (2008). Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on intelligent user interfaces* (pp. 227–236). Assocation for Computing Machinery.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.

Govier, T. (1994). Is it a jungle out there? Trust, distrust and the construction of social reality. *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie*, 33(2), 237–252.

Guo, L., Daly, E. M., Alkan, O., Mattetti, M., Cornec, O., & Knijnenburg, B. (2022). Building trust in interactive machine learning via user contributed interpretable rules. In *Proceedings of the 27th international conference on intelligent user interfaces* (pp. 537–548). Assocation for Computing Machinery.

Guzman, A. (2018). What is human-machine communication, anyway? In A. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 1–29). Peter Lang.

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, 22(1), 70–86.

Hall, M. A. (2002). Law, medicine, and trust. *Stanford Law Review*, 55(2), 463–527.

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

Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation.

Hardré, P. L. (2016). When, how, and why do we trust technology too much? In *Emotions, technology, and behaviors* (pp. 85–106). Academic Press.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105, 105–120.

Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., & Martín, N. (2021). *How humans judge machines*. MIT Press.

High-Level Expert Group (HLEG). (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hill, C. A., & O'Hara, E. A. (2006). A cognitive theory of trust. *Washington University Law Review*, 84, 1717–1796.

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

Hult, D. (2018). Creating trust by means of legislation–a conceptual analysis and critical discussion. *The Theory and Practice of Legislation*, 6(1), 1–23.

Im, H., Sung, B., Lee, G., & Kok, K. Q. X. (2023). Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude. *Computers in Human Behavior*, 145, 107791.

Itoh, M. (2012). Toward overtrust-free advanced driver assistance systems. *Cognition, Technology & Works*, 12, 51–60.

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.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR)*, 55(2), 1–38.

Keymolen, E. (2016). *Trust on the line: A philosophical exploration of trust in the networked era*. Wolf Legal Publishers.

Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 262–271). Assocation for Computing Machinery.

Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, 35(2), 1–20.

Kundinger, T., Wintersberger, P., & Riener, A. (2019). (Over) Trust in automated driving: The sleeping pill of tomorrow? In *Extended abstracts of the 2019 CHI conference on human factors in computing systems* (pp. 1–6). Assocation for Computing Machinery.

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

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

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

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

Long, C. P., & Sitkin, S. B. (2018). Control–trust dynamics in organizations: Identifying shared perspectives and charting conceptual fault lines. *Academy of Management Annals*, 12(2), 725–751.

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

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

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11th Australasian conference on information systems*. Association for Information Systems.

Malone, M. F. T. (2010). The verdict is in: The impact of crime on public trust in Central American justice systems. *Journal of Politics in Latin America*, 2(3), 99–128.

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

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, 53(4), 356–370.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, 55(3), 520–534.

Meyerson, D., Weick, K. E., & Kramer, R. M. (1996). Swift trust and temporary groups. *Trust in organizations: Frontiers of Theory and Research*, 166, 195.

Möllering, G. (2006). *Trust: Reason, routine, reflexivity*. Emerald Group Publishing.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, 38(2), 311–322.

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

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human–Computer Studies*, 45, 669–678.

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology*, 29, 1093–1110.

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

Newlands, G., Lutz, C., Tamò-Larrieux, A., Villaronga, E. F., Harasgama, R., & Scheitlin, G. (2020). Innovation under pressure: Implications for data privacy during the Covid-19 pandemic. *Big Data & Society*, 7(2), 205395172097668.

Nielsen, J., Clemmensen, T., & Yssing, C. (2002). Getting access to what goes on in people's heads? Reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on human-computer interaction* (pp. 101–110). Assocation for Computing Machinery.

NIST. (2021). Trust and artificial intelligence. https://www.nist.gov/publications/trust-and-artificial-intelligence

Nix, J. (2017). Police perceptions of their external legitimacy in high and low crime areas of the community. *Crime & Delinquency*, 63(10), 1250–1278.

Nooteboom, B., & Six, F. (2003). The trust process. In *The trust process in organizations: Empirical studies of the determinants and the process of trust development* (pp. 16–36). Edward Elgar Publishing.

Nowotny, H. (2021). *In AI we trust: Power, illusion and control of predictive algorithms*. John Wiley & Sons.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410.

Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6), 542–556.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, 49(1), 95–112.

Ribstein, L. E. (2001). Law v. trust. *Boston University Law Review*, 81, 553.

Richards, N., & Hartzog, W. (2021). A duty of loyalty for privacy law. *Washington University Law Review*, 99, 961–1021.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In *Proceedings of the 11th annual ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 101–108). Assocation for Computing Machinery and the Institute of Electrical and Electronics Engineers.

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.

Russell, S. (2020). *Human-compatible artificial intelligence*. Penguin LLC US.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) Robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the 10th annual ACM/IEEE international conference on human-robot interaction* (pp. 141–148). Assocation for Computing Machinery and the Institute of Electrical and Electronics Engineers. https://doi.org/10.1145/2696454.2696497

Sanders, T. L., Wixon, T., Schafer, K. E., Chen, J. Y. C., & Hancock, P. A. (2014). The influence of modality and transparency on trust in human-robot interaction. In *Proceedings of the IEEE international inter-disciplinary conference on cognitive methods in situation awareness and decision support (CogSIMA)* (pp. 156–159). IEEE.

Schaefer, K. (2013). The perception and measurement of human-robot trust. University of Central Floridae, Electronic Thesis and Dissertations. https://stars.library.ucf.edu/etd/2688/

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.

Searle, R., Weibel, A., & Den Hartog, D. N. (2011). Employee trust in organizational contexts, Chapter 5. In G. P. Hodgkinson & J. K. Ford (Eds.), *International review of industrial and organizational psychology, 2011* (Vol. 26). John Wiley & Sons, Ltd.

Shao, Z., Li, X., Guo, Y., & Zhang, L. (2020). Influence of service quality in sharing economy: Understanding customers' continuance intention of bicycle sharing. *Electronic Commerce Research and Applications*, 40, 100944.

Shapiro, S. J. (2011). *Legality*. Harvard University Press.

Shapiro, S. P. (1987). The social control of impersonal trust. *American Journal of Sociology*, 93(3), 623–658.

Sharma, S. (2023). EU closes in on AI act with last-minute ChatGPT-related adjustments. *Computerworld*. https://www.computerworld.com/article/3695009/eu-closes-in-on-ai-act-with-last-minute-chatgpt-related-adjustments.html

Sheppard, B. H., & Sherman, D. M. (1998). The grammars of trust: A model and general implications. *Academy of Management Review*, 23(3), 422–437.

Starke, G., & Ienca, M. (2022). Misplaced trust and distrust: How not to engage with medical artificial intelligence. *Cambridge Quarterly of Healthcare Ethics*, 1–10. https://doi.org/10.1017/S0963180122000445.

Strauss, A., & Corbin, J. (1994). Grounded theory methodology: An overview. In N. Denzin & Y. Lincoln (Eds.), *Handbook of Qualitative Research* (1st ed., pp. 273–284). SAGE Publications.

Sztompka, P. (1999). *Trust: A sociological theory*. Cambridge University Press.

Tamò-Larrieux, A. (2021). Decision-making by machines: Is the 'Law of Everything' enough? *Computer Law & Security Review*, 41, 105541.

Thaler, R. H. (2000). From homo economicus to homo sapiens. *Journal of Economic Perspectives*, 14(1), 133–141.

The Economist (2021). China rapidly shifts from a two-child to a three-child policy, 3 June. https://www.economist.com/china/2021/06/03/china-rapidly-shifts-from-a-two-child-to-a-three-child-policy

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Vestager, M. (2021). *Trust and technology in a new digital age*. European Commission https://ec.europa.eu/commission/commissioners/2019-2024/vestager/announcements/trust-and-technology-new-digital-age_en

Wan, E. W., & Chen, R. P. (2021). Anthropomorphism and object attachment. *Current Opinion in Psychology*, 39, 88–93.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *In 2016 11th annual ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 109–116). IEEE.

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

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, 2, 352–367.

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *International Journal of Social Robotics*, 4(3), 235–248.

Yamada, K., & Kuchar, J. K. (2006). Preliminary study of behavioral and safety effects of driver dependence on a warning system in a driving simulator. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 36(3), 602–610.

Yang, R., & Wibowo, S. (2022). User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets*, 32, 2033–2057. https://doi.org/10.1007/s12525-022-00592-6

Yeung, J. (2023). China's population is shrinking. The impact will be felt around the world. CNN.

Zucker, L. G. (1986). Production of trust: Institutional sources of economic structure, 1840–1920. *Research in Organizational Behavior*, 8, 53–111.